Research White Paper

Mechanistic Evidence in Evidence-Based Medicine: A Conceptual Framework

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

AHRQ

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Research White Paper

Mechanistic Evidence in Evidence-Based Medicine: A Conceptual Framework

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No: 290-2007-10061-I

Prepared by:
The Johns Hopkins University Evidence-based Practice Center
Baltimore, MD

Investigators:
Steven N. Goodman, M.D., Ph.D.
Jason Gerson, Ph.D.

AHRQ Publication No. 13-EHC042-EF
June 2013

This report is based on research conducted by the The Johns Hopkins University Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10061-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Goodman SN, Gerson J. Mechanistic Evidence in Evidence-Based Medicine: A Conceptual Framework. Research White Paper (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No. 290-2007-10061-I). AHRQ Publication No. 13-EHC042-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Research White Paper. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Task Order Officer, Director, Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

Karen Robinson, M.Sc., Ph.D.
Johns Hopkins University

Harry Marks*, Ph.D.
Johns Hopkins University

*Deceased, 2010

Mechanistic Evidence in Evidence-Based Medicine: A Conceptual Framework

Structured Abstract

Background. Virtually all current frameworks for the evaluation of the strength of evidence for an intervention's effect focus on the quality of the design linking the intervention to a given outcome. Knowledge of biological mechanism plays no formal role. In none of the evidence grading schemas, new statistical methodologies or other technology assessment guidelines is there a formal language and structure for how knowledge of how an intervention works.

Objectives. The objective was to identify and pilot test a framework for the evaluation of the evidential weight of mechanistic knowledge in evidence-based medicine and technology assessment.

Methods. Six steps were used to develop a framework for the evaluation of the evidential weight of mechanistic knowledge: (1) Focused literature review, (2) Development of draft framework, (3) Workshop with technical experts, (4) Refinement of framework, (5) Development of two case studies, (6) Pilot test of framework on case studies.

Results. The final version of the framework for evaluation of mechanistic evidence incorporates an evaluation of the strength of evidence for the:

1. Intervention's target effect in nonhuman models.
2. Clinical impact of target effect in nonhuman models.
3. Predictive power of nonhuman model for an effect in humans
   3t. The predictive power of the target effect model
   3c. The predictive power of the clinical effect model
4. Intervention's target effect in human disease states.
5. Clinical impact of the target effect in human disease states.

A graphic representation is included in the full report.

Conclusion. This framework has several features combining work from a variety of fields that represent an important step forward in the rigorous assessment of such evidence.

1. It uses a definition of evidence based on inferential effect, not study design.
2. It separates evidence based on mechanistic knowledge from that based on direct evidence linking the intervention to a given clinical outcome.
3. It represents the minimum sufficient set of steps for building an indirect chain of mechanistic evidence.
4. It is adaptable and generalizable to all forms of interventions and health outcomes.

It mirrors in the evidential framework the conceptual framework for translational medicine, thus linking the fields of basic science, evidence-based medicine and comparative effectiveness research.

Contents

Figures

Appendixes

Executive Summary

Background

Virtually all current frameworks for the evaluation of the strength of evidence for an intervention's effect focus on the quality of the design linking the intervention to a given outcome. Knowledge of biological mechanism plays no formal role, in spite of the fact that such knowledge is typically the basis for the development of the intervention. At best, mechanistic knowledge comes in indirectly, through the choice of endpoints, target populations, and perhaps under the vague rubric of "biological plausibility." But nowhere in any of the evidence grading schemas, new statistical methodologies or other technology assessment guidelines do we have a formal language and structure for how knowledge of how an intervention works should enter the process.

Objective

Our objective was to identify and pilot test a framework for the evaluation of the evidential weight of mechanistic knowledge in evidence-based medicine and technology assessment.

Methods

We used multiple resources and perspectives to help us develop a framework the evaluation of the evidential weight of mechanistic knowledge. We carried out the following six steps:

Step 1—Focused literature review
Step 2—Development of draft framework
Step 3—Workshop with technical experts
Step 4—Refinement of framework
Step 5—Development of two case studies
Step 6—Pilot test of framework on case studies

Results

Step 1—Focused Literature Review

We conducted comprehensive literature reviews in two broad areas: evaluation of surrogate endpoints and the value and use of animal models in translational research. Both searches encompassed the publication dates between 2000 and 2009, with additional references found before and after those dates through reference and citation searches. Reviews were conducted on 125 articles on animal models, 133 on surrogate markers, and 24 on evidential grading systems. An annotated bibliography summarized 93 of the articles on animal models as well as 103 of the articles on certain points. All of these articles were mapped into a preliminary version of a conceptual framework.

Step 2– Development of Draft Framework

Based on preliminary review of related literature, an initial draft framework was devised and used as the basis for mapping the annotated bibliographies, as well as discussion in a subsequent workshop. This initial framework was substantively modified on the basis of the work conducted for this report and is presented below:

Strength of evidence for:

1. Existence of pathway.
2. Existence of pathway in humans.
3. Completeness of pathway.
4. Alternate, competing, or compensatory pathways
5. Similarity to other interventions/mechanisms with known clinical effects
6. Adverse event mechanisms

Step 3—Invited Workshop With Technical Experts

An exploratory workshop was held with experts in translational medicine, toxicology, philosophy evidence-based medicine and a variety of other fields. A draft conceptual framework was presented and discussed, and each participant presented their own experience and knowledge concerning the use of mechanistic knowledge in either interpreting or developing research on emerging therapies. The workshop was summarized and the conceptual framework revised.

Step 4—Development of Framework

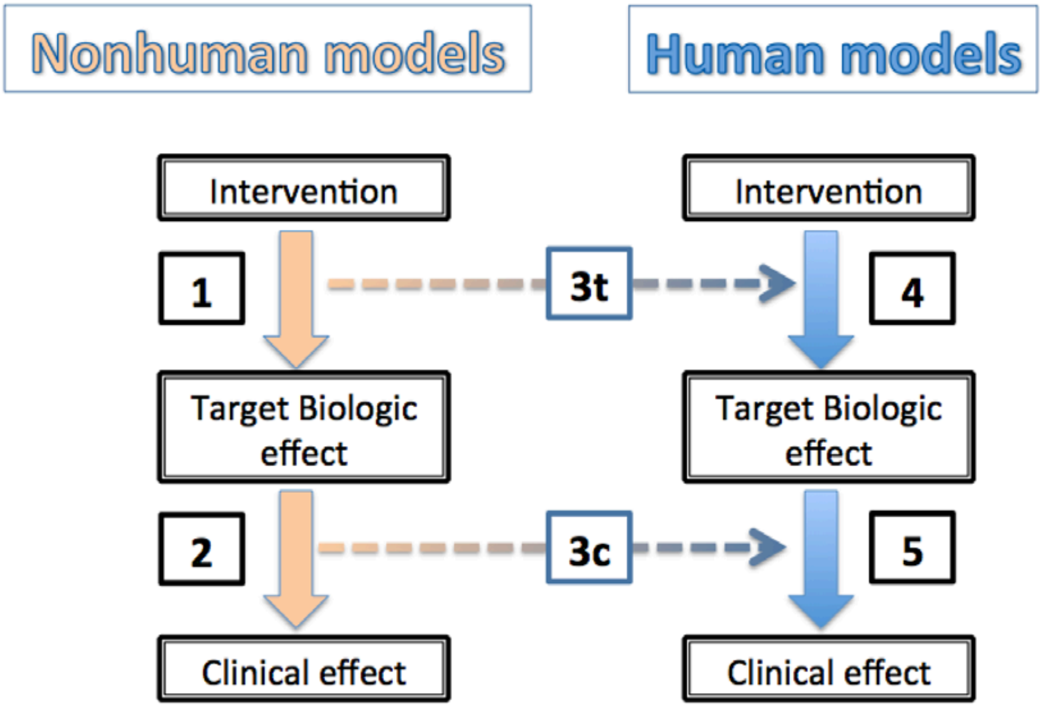
Based our review of the literature described above as well as the technical input gathered in the workshop on the draft framework, we propose the following final version of the framework for evaluation of mechanistic evidence:

Strength of evidence for the:

1. Intervention’s target effect in nonhuman models.
2. Clinical impact of target effect in nonhuman models.
3. Predictive power of nonhuman model for an effect in humans
   - 3t. The predictive power of the target effect model
   - 3c. The predictive power of the clinical effect model
4. Intervention’s target effect in human disease states.
5. Clinical impact of the target effect in human disease states.

This is the minimally sufficient series of steps necessary for such a framework, and this has sufficient generality to apply to virtually all types of interventions. This framework is demonstrated graphically in Figure A below:

Figure A. Schematic of mechanistic framework model



The propagation of the strength of evidence is through a Bayesian algorithm, with the strength of evidence represented by the degree to which the probability of a clinical effect is modified by evidence from the component steps. This modeling makes clear how strong mechanistic evidence can be necessary for proper inferences, yet still, by itself, yield very low probabilities of success for a given intervention.

Step 5—Pilot Test of Framework in two Case Studies

As part of a companion project, two in depth case studies were developed to see how the conceptual framework being developed would apply to actual examples. The two case studies were of Gleevec for the treatment of chronic myelogenous leukemia and estrogen use in menopausal women for the prevention of heart disease. These case studies were summarized and mapped into the conceptual framework.

Discussion

We utilized multiple resources and perspectives including literature review and consultation with experts at our institution to develop a framework for the use of mechanistic knowledge in the evaluation of the effectiveness of medical interventions. This framework has several features combining work from a variety of fields that represent an important step forward in the rigorous assessment of such evidence.

1. It uses a definition of evidence based on inferential effect, not study design.
2. It separates evidence based on mechanistic knowledge from that based on direct evidence linking the intervention to a given clinical outcome.
3. It represents the minimum sufficient set of steps for building an indirect chain of mechanistic evidence.

4. It is adaptable and generalizable to all forms of interventions and health outcomes.
5. It mirrors in the evidential framework the conceptual framework for translational medicine, thus linking the fields of basic science, evidence-based medicine and comparative effectiveness research.

Limitations and Future Research

While we believe the framework provided to be the starting point for any discussions of the value of mechanistic knowledge, much remains to be done in the form of both further refinement and implementation. In terms of refinement, while the framework components themselves represent a minimally sufficient set of dimensions, the optimal set of component questions within each of these dimensions requires further work. The more specificity that is provided in the sub questions, the more operational the framework becomes, but also potentially the more limited.

More work must also be done on how best to quantitate or weigh the impact both within and between various dimensions. Because many of the inferences cannot fall back on randomization, the same kinds of evidential judgments used when assessing observational studies must be applied to many of these designs. Building such a quantitative network or chain of inferences are similar to complex quantitative risk models, and the relevance of such techniques to this application should be explored. As shown in some of the examples provided, it is possible to roughly quantitate the evidential value of the entire drug development process; refining this for specific interventions or in nondrug applications, will require substantially more work, yet is clearly achievable.

The pilot examples of the use of the framework demonstrated both its potential strength and areas for further work. It was clear in both cases that the framework could be applied, and that such application could illuminate those domains in which the evidence made the relationship between the therapy and the outcome more or less likely. In both cases, we saw that a limited number of pathways, a well characterized pathophysiology, accurate measures and a clearly delineated target within those pathways were key elements. However, how the various qualitative observations can be quantitatively assessed and the relative weights of various dimensions, or algorithmic combination thereof, requires further work.

Conclusions

The formal language and logic of evidential assessment in evidence-based medicine and comparative effectiveness research has no formal place for incorporating knowledge of “how things work” in medicine. This project has provided a conceptual framework for that assessment, with proposals for how this might be combined with direct evidence to provide a way of capturing all the ways of knowing in medicine, defined both on the group level and at the level of the individual.

Much work remains to be done in terms of refining the subcomponents of these dimensions and in their quantification or combination. Further developing this framework can help not only in the accurate representation of evidence for therapeutic decisionmaking and medical policy, but can potentially speed the development of medical interventions by demonstrating how and where mechanistic evidence can augment direct evidence. A potentially even more important outcome is that this framework can help bring together those communities working on the development and the assessment of therapies, who rarely seem to communicate except occasionally at the translational divide, and whose different views of what constitutes legitimate evidence has been a source of both misunderstanding and indeed conflict between those communities of researchers.

Developing a common framework for evidence may be a first step towards true interdisciplinary, translational knowledge

Background

Interest has increased in recent years in “comparative effectiveness,” that is, assessing the efficacy of new or established medical interventions, with particular emphasis on head to head comparisons of established therapies, or understanding their real-world performance. Randomized controlled trials (RCTs), while the ostensible gold standard for establishing efficacy and sometimes effectiveness, have well recognized liabilities, most notably the time and expense it often takes to mount them, as well as the sometimes limited scope of the questions they address.

Alternatives to RCTs include a variety of observational designs. Those attracting considerable attention are typically derived from very large databases, often assembled for non-research purposes, such as hospital billing, reimbursement, prescription data, electronic patient records, etc. Studies derived from such data sources promise real-world relevance, and relatively rapid results, compared to some RCTs. The middle ground is occupied by observational designs with original data gathering and RCTs that utilize surrogate endpoints—for example, death versus tumor progression, elevated LDL or coronary artery narrowing versus MI or death.

A dilemma facing patients, physicians, regulatory entities, insurance providers, guideline developers and others with an interest in evidence assessment involves: (1) how pertinent existing RCT evidence is to the decisions they have to make and (2) how informative and reliable results from either observational designs or RCTs that use surrogate outcomes are in determining either efficacy or effectiveness. It is generally recognized that observational designs are subject to subtle biases that can have large effects (e.g., WHI), and that data not gathered for research purposes often lacks the precision or validity to make reliable inferences. The main approaches to these problems currently being discussed are three-fold; improving the quality and completeness of the underlying data, using innovative statistical methodologies to diminish the effects of confounding, and the development of evidence grading schemes to distinguish reliable from unreliable evidence. The ultimate goal of such efforts is to derive conclusions through these approaches that are nearly as reliable and perhaps more relevant for policy purposes than RCTs.

What is notably absent from these conversations is the role that should be played by knowledge of mechanism, and how this can help in the evaluation of observational evidence, including the detection of effect modification (e.g., “personalized medicine”). With the ascendance of the evidence-based medicine, there is no formal role for mechanistic knowledge in the evidence-evaluation framework. At best, mechanistic knowledge comes in indirectly, through the choice of endpoints, target populations, and perhaps under the vague rubric of “biological plausibility.” But nowhere in any of the evidence grading schemas, new statistical methodologies or other technology assessment guidelines do we have a formal language and structure for how knowledge of how an intervention works should enter the process. The closest we have is in the prior probability distribution functions of Bayesian approaches, but this begs the question of how to reliably determine how much mechanistic knowledge is worth.

Defining "Evidence"

Evidence-based medicine defines the strength of evidence in terms of how the information was produced, e.g., RCT, case-control study, case series, etc. (Harris *et al.*, 2001; Atkins *et al.*, 2004) The definition to be employed here is based instead on its inferential effect, following principles of Bayesian inference and probabilistic causality (Suppes, 1970). We define evidence as information that modifies the probability that an intervention will have a non-null causal effect on an outcome. The strength of evidence is related to the magnitude of change in that probability, with the most familiar Bayesian metric being the Bayes factor or its logarithm. (Good, 1950; Kass and Raftery, 1995; Royall, 1997; Goodman, 1999). It is expressed in a simple version of Bayes theorem as follows:

Bayes Theorem

Prior odds of clinical effect x Bayes Factor = Posterior odds of a clinical effect

$$\text{Bayes factor } (H_A \textit{ vs. } H_0) = \frac{\text{Probability of observed data under } H_A}{\text{Probability of observed data under } H_0}$$

Where:
H_0 = Null hypothesis that intervention has no clinical effect
H_A = Alternative hypothesis that intervention has a clinical effect

In the case of diagnostic tests, the Bayes factor is equivalent to the likelihood ratio commonly used in EBM (Good, 1950; Kass and Raftery, 1995; Royall, 1997). As in the case of diagnostic testing, it is critical to separate the posterior probability of a hypothesis (aka, the predictive value of a given result) from the strength of evidence. A diagnostic test can be extremely powerful, yet if the prevalence of the tested disease (a.k.a. prior probability of disease) is low enough, the positive predictive value of that test can still be quite low. Similarly, mechanistic evidence for an intervention's effect can be very powerful without making the probability of that effectiveness high. For the probability of effectiveness to be high, evidence from clinical research is usually required; how much is in part determined by the strength of the mechanistic evidence. As was noted in a description of the role of mechanistic evidence in the IARC determinations of carcinogenicity, "There is an implicit trade-off between the strength of the evidence in humans and the strength of the mechanistic data needed: the weaker the evidence in humans, the stronger the mechanistic data must be to warrant a classification [as a human carcinogen]."

Under the theory of probabilistic causality, a cause is defined as condition whose presence or absence, all other factors being equal, changes the probability of an outcome. (Suppes, 1970) Thus, all interventions with non-zero clinical effects, independent of other factors, are by definition causes of the outcome. This links the evidential criteria used in EBM and proposed herein to causal criteria proposed in epidemiology and clinical medicine (Hill, 1965; Susser, 1977). We will therefore use the language and framework of causation interchangeably with that of therapeutic effectiveness.

The Strength of Evidence Provided by the Preclinical Development Process

The Bayes factor can be used to roughly quantify the strength of evidence provided by pre-clinical drug development. Many in the pharmaceutical industry lament the poor yield of pre-clinical drug testing, but this low yield is a reflection of the posterior probability of success, not the evidential value of the research. Based on data from pharmaceutical companies from 1991-2000, Kola (Kola and Landis, 2004) reported that roughly 10 percent of drugs entering Phase I studies were ultimately approved for use, with about half of failures due to nonclinical reasons (e.g., market considerations). It has been estimated that roughly 1000 compounds are screened for every one that enters clinical testing. Thus, the ratio of screened compounds to approved therapies is about 10,000:1, for a yield of about 1/10,000 if we picked a compound at random to test clinically. While the success rate of the pre-clinical process seems low (10 percent), it has increased 1000 times over the success rate to be expected by choosing compounds at random. So the pre-clinical process is hugely informative, raising the odds of success about 1000-fold, with that multiplier being the value of the Bayes factor.

In contrast, if we presume that the clinical development process must raise the probability of clinical efficacy from 10 percent to 95 percent, that requires a Bayes factor of only about (95/5)÷(10/90) = 171, and about half that level if nonclinical failures are omitted. This informal calculation shows that while pre-clinical evidence may provide an insufficient basis to upon which choose human therapies, it still provides quantitatively more evidence than the clinical phase of testing, which in this example provided only 43 percent of the total (=log 171/(log 1000 + log 171)) . Thus, mechanistic information has substantial evidential value, and its strength affects the amount of evidence required from the clinical testing process. It is critical to note how different that perspective is from that of EBM, which relegates such preclinical work to the realm of "nonevidence". The usage here is consistent with that of Vandenbroucke, who notes that the predictive power of stronger designs may be due more to the prior probability needed before such a design is implemented than the evidential strength of the design itself. (Vandenbroucke, 2004)

In another study, Contopoulos et al. (Contopoulos-Ioannidis *et al.*, 2003) looked at 25,190 basic science studies published around 1980 and found 101 in which there were claims of potential clinical utility. Of these, 19 resulted in at least one positive clinical trial, and 5 were ultimately licensed for clinical use. Using the latter number as a count of the number of successful technologies, we see numbers not qualitatively dissimilar from those reported for drug development; approximately 250 basic science studies per technology that entered clinical development (25190/101), with 1 in 20 of these (5/101) ultimately succeeding. Again, *the basic research process has increased the odds of success more than does the clinical research process*, this time by a factor of about 12 (250/20). So while the translational and developmental process is often decried for its small yield of usable therapies, it must be recognized that it increases the odds of success by several orders of magnitude, leaving the clinical evaluation process to increase it yet further to justify clinical use.

Objective

Our objective was to identify and pilot test a framework for the evaluation of evidence from knowledge of biological mechanism.

Methods

We used multiple resources and sought different perspectives to develop a framework for the identification of research gaps. We carried out six steps. We first attempted to identify, enumerate and describe frameworks that have been used (Steps 1 to 3). We then developed, tested and refined a framework (Steps 4 to 6). The six steps are:

Step 1—Focused literature review
Step 2—Development of draft framework
Step 3—Workshop with technical experts
Step 4—Refinement of framework
Step 5—Development of two case studies
Step 6—Pilot test of framework on case studies

Step 1—Focused Literature Review

We conducted comprehensive literature reviews in two broad areas: evaluation of surrogate endpoints and the value and use of animal models in translational research. Both searches encompassed the publication dates between 2000 and 2009, with additional references found before and after those dates through reference and citation searches. Criteria for inclusion in this report was the degree to which the paper provided either a high-level perspective or a case example directly relevant to the development of a framework. These two searches were used to compile an annotated bibliography (Appendices A and B). From each included article we also mapped its focus into the draft framework we were using at that time.

Step 2—Development of Draft Framework

On the basis of preliminary reading in each of the domains described previously, the following draft framework was proposed. This was both discussed at the invited workshop, and used for the mapping exercise in the annotated bibliography.

Step 3—Workshop With Technical Experts

We identified technical experts representing a variety of disciplines relevant to the development of this framework. These included translational medicine, biomarker development, philosophy, evidence-based medicine, toxicology and animal research. The proceedings of the workshop are presented in Appendix C.

The discussion in the workshop covered a very broad territory, reflecting the range of expertise among the participants. Each participant presented an example from their own domain of work of the use of mechanistic information in the development, evaluation, or prediction of the efficacy of a therapy. These perspectives then informed conversation focused specifically on the components of the framework, as well as providing potential examples for its application. The discussions in this workshop were quite rich, going far beyond the issue of the framework itself. However, both the examples and ensuing discussion of the framework highlighted the following potential weaknesses in the draft framework as proposed:

1. Virtually all pathways underlying biologic mechanisms are incompletely known.
2. Most therapies are developed based on partial knowledge of mechanisms, and rarely affect more than one step in a pathway.
3. The quality and relevance of experiments in animals to human disease and therapeutics are widely recognized as deeply problematic.

4. Experimentation on animals serves many purposes other than documentation of whole organism responses. The use of transgenic organisms, and other genetically manipulated animals makes them occasionally excellent experimental models for specific therapeutic effects.

As a result of these observations plus other discussion in the workshop along with further literature searching the originally proposed framework was modified as described in the next section.

Step 4—Refinement of the Framework

To be maximally useful, an evidential framework for mechanistic knowledge must be applicable to all forms of interventions in humans to prevent or treat disease. That requires a high degree of generality for the overall structure, with elements that are customizable for particularly contexts, e.g., drugs, devices or behavioral interventions. Second, it to should have the irreducible minimum of elements, capturing only those that are absolutely essential for the task. Finally, the potential application and context in which this framework is expected to be used must be clear. The main domains of application are in those settings where empirical information directly linking intervention to human outcome is absent or weak, such as typically occurs in early phase clinical testing or technological development, but also arises in many other contexts, listed in Table 1.

These considerations lead to the development of an alternative framework which included many of the ideas embedded in the draft framework, but organize them in a way that reflected the developmental processes of therapeutics as well as the minimally sufficient set of conditions and categories for evidential measurement.

Step 5—Pilot Test of the Framework on Two Case Studies

As part of a companion project, two in depth case studies were developed to see how the conceptual framework being developed would apply to actual examples. The two case studies were of Gleevec for the treatment of chronic myelogenous leukemia and estrogen use in menopausal women for the prevention of heart disease. These case studies were summarized and mapped into the conceptual framework.

Results and Case Studies

Step 1—Focused Literature Review

We conducted comprehensive literature reviews in two broad areas: evaluation of surrogate endpoints and the value and use of animal models in translational research. Both searches encompass the publication dates between 2000 and 2009, with additional references found before and after those dates through reference and citation searches. 125 articles on animal models were reviewed, 133 on surrogate markers and 24 on evidential grading systems. 93 of the articles on ‘ animal models were summarized in an annotated bibliography, as were 103 of the articles on certain points. All of these articles were mapped into a preliminary version of a conceptual framework because the modified framework had not been developed when this was originally done. These bibliographies can be found in Appendices A and B.

Step 2—Development of Draft Framework

On the basis of preliminary reading in each of the domains described previously, the following draft framework was proposed. This was both discussed at the invited workshop, and used for the mapping exercise in the annotated bibliography. This draft framework was as follows:

Strength of evidence for:

1. Existence of pathway.
2. Existence of pathway in humans.
3. Completeness of pathway.
4. Alternate, competing or compensatory pathways
5. Similarity to other interventions / mechanisms with known clinical effects
6. Adverse event mechanisms

Step 3—Workshop With Technical Experts

We identified technical experts representing a variety of disciplines relevant to the development of this framework. These included translational medicine, biomarker development, philosophy, evidence-based medicine, toxicology and animal research. The proceedings of the workshop are presented in Appendix C.

The discussion in the workshop covered a very broad territory, reflecting the range of expertise among the participants. Each participant presented an example from their own domain of work of the use of mechanistic information in the development, evaluation, or prediction of the efficacy of a therapy. These perspectives then informed a conversation focused specifically on the components of the framework, as well as providing potential examples for its application. The discussions in this workshop were quite rich, going far beyond the issue of the framework itself. However, both the examples and ensuing discussion of the framework highlighted the following potential weaknesses in the draft framework as proposed:

1. Virtually all pathways represent biologic mechanisms that are incompletely known.
2. Most therapies are developed based on partial knowledge of mechanisms, and attempt to affect just one step in a pathway, typically called a “target”.
3. Most clinical effects are mediated through multiple pathways.

4. Both the quality and relevance to humans of experiments in animals are widely recognized as deeply problematic, yet still important to refine or winnow the list of candidate interventions.
5. Experimentation in animals serves many purposes other than documentation of whole organism responses. The use of transgenic organisms, and other genetically manipulated animals makes them occasionally excellent experimental models for specific therapeutic effects.

As a result of these observations plus discussion in the workshop along with further literature searching, the originally proposed framework was modified as described in the next section.

Step 4—Development of Framework

To be maximally useful, an evidential framework for mechanistic knowledge must be applicable to all forms of interventions in humans to prevent or treat disease. That requires a high degree of generality for the overall structure, with elements that are customizable for particularly contexts, e.g., drugs, devices or behavioral interventions. Second, it should have the irreducible minimum of elements, capturing only those that are absolutely essential for the task. Finally, the potential application and context in which this framework is expected to be used must be clear. The main domains of application are in those settings where empirical information directly linking intervention to human outcome is absent or weak, such as typically occurs in early phase clinical testing or technological development, but also arises in many other contexts, listed in Table 1.

The main approach and the revision of the framework was therefore to consider what were the minimal requirements for what could be called a biologic mechanism. The existence of a “target,” a single intermediate step in the causal pathway between the intervention and the outcome, is central to the definition of a biologic mechanism. The existence of such a step is a minimal, necessary condition for mechanistic knowledge to be informative about the intervention-outcome relationship. Without this step, the totality of the evidence is from the empirically observed relationship between intervention and outcome, which by definition is not mechanistic. Therefore, this framework does not include linkages that directly connect an intervention and an outcome, which represent evidence that is important, but it is not “mechanistic”. A target is defined here as a necessary step in a sufficient pathway, or a component of a sufficient cause.

The use of the target connects this framework to the development of therapeutic interventions, wherein a molecular or mechanistic target for an intervention is identified. This target can take a physical form, such as a protein binding site or metabolic process, or it could be a psychological or even sociological state that is part of a proposed mechanism. Any component of a mechanism can in theory be a target for intervention.

These considerations lead to the development of an alternative framework that reflected the minimally sufficient set of conditions and categories for evidential measurement and perhaps not coincidentally, the developmental processes of therapeutics as well. The language of “pathways” in the original draft framework was too limiting, in that it is rare that all steps in a given pathway are understood, and even when they are, such as in the coagulation cascade, typically only one is targeted with a particular agent. Thus the concept of mechanism must involve at least one intermediate step, with an attendant theory about how that step (here called the target) is part of the causal chain. The revised framework was as follows:

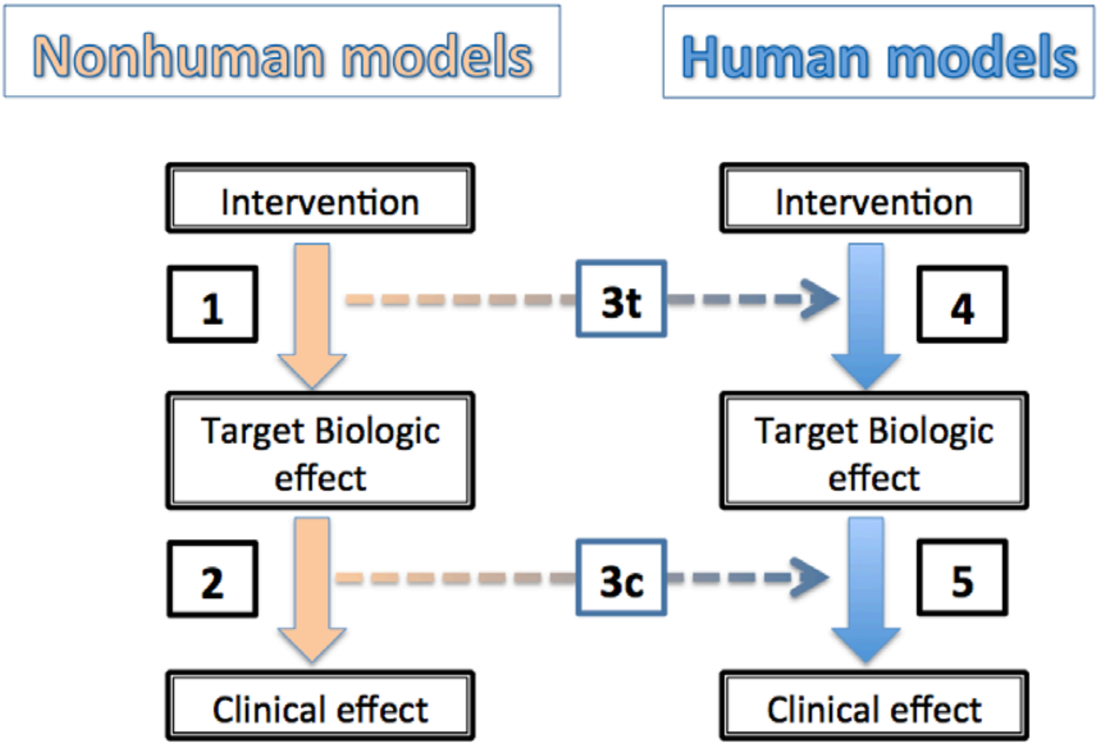
Strength of evidence for the:

1. Intervention’s target effect in nonhuman models.
2. Clinical impact of target effect in nonhuman models.

3. Predictive power of nonhuman model for an effect in humans
   3t. The predictive power of the target effect model
   3c. The predictive power of the clinical effect model
4. Intervention's target effect in human disease states.
5. Clinical impact of the target effect in human disease states.

This framework is demonstrated graphically in Figure 1 below:

Figure 1. Schematic of mechanistic framework model



This is the minimally sufficient series of steps necessary for such a framework, and has sufficient generality to apply to virtually all types of interventions. It represents a "chain of evidence" that is often used in other evidential frameworks and risk assessment models to assess cumulative evidential strength. This shows how the evidence in humans trumps that in nonhuman models, in that the nonhuman evidence requires a "translation" step (3t or 3c), so that the stronger the human evidence, the less necessary that from the indirect sources. The weaker the human evidence, the more the linkages from 1, 2, and 3 are critical to the inference. Each one of these dimensions has a set of distinctive characteristics that need to be assessed, which will be discussed below.

Intervention's Target Effect in Nonhuman Models

All interventions and biologic processes work though at least one intermediate target step that is necessary but not sufficient to produce a biologic effect. A series of such steps, mechanistically linked, is called a "pathway." Most therapies act on, or are designed to act on, just one key step in that pathway through a "target". Anticoagulants typically interfere with the coagulation cascade at an identifiable point or points. If a proposed anticoagulant had no demonstrable effect on any step in that cascade, this would serve as strong mechanistic evidence against a clinical effect. If the key target step is purely theoretical, never having been demonstrated in a biologic system, any claim based on such a mechanism would have negligible weight. Occasionally, this step is not consistent with known physical principles. This might include homeopathic remedies with no detectable agent in the preparation, a claimed radiation effect when it can be demonstrated that the target tissue receives no exposure or a vaccine that is not antigenic.

This key step often defines the class of intervention of drug, for example, calcium channel blockers, proton-pump inhibitor, tyrosine kinase inhibitors, cardiac bypass procedure, etc. While other details of mechanism typically differ somewhat among interventions in a class, they all have some form of the key step in common, with either the identical target effect, or effect on the same pathway in which the target exists. The key elements of evidential quality related to this step are as follows:

1. Evidence that it is physically possible for the intervention to exert a physical effect through the target (e.g., absorption, PK, PD for drugs, ability to place a cardiac stent in a narrowed artery).
2. Evidence for the validity of the measure of the effect on the target. (e.g., that the assay that confirms receptor site blockage is an accurate measure of that blockage.)
3. Evidence that intervention in question has desired impact on the measure of target effect.

Embedded within each of the above evidentiary requirements are a set of components that are intervention or target specific, including components of good study design, and research on related interventions and targets. It includes both the qualitative considerations and the quantitative strength of relationships. The evidential framework for biomarker validation described by Altar (Altar et al., 2008; Altar, 2008) contains most of the above elements, including the explicit mention of the existence of a theory, which is presupposed here.

Clinical Impact of Target Effect in Nonhuman Models

This next step in the evidence framework involves linking the effect on the target to some whole-organism impact that parallels a human clinical outcome. The strength of evidence for this step includes:

- Evidence for the analytic validity of the clinical impact measure. (It must be recognized that it increases the odds of success by several orders of magnitude, leaving the clinical evaluation process to increase it yet further.)
- Evidence relating target effect to measure of clinical impact.

This dimension concerns the relationship between the target effect and the clinical outcome. This can include evidence both with and without the intervention. For example, if there are ways to induce the target effect without the intervention, this can serve as supporting evidence for this step. Similarly, among groups who experienced intervention, supporting evidence would be a differential clinical effect among those who had a demonstrable target effects versus those who had not. For example, if we were examining an agent that improve surgical mortality by decreasing operative blood loss, supportive evidence for this step could include experiments showing such improved mortality when blood loss was reduced by means other than the agent under study. If, on the other hand, all subjects were given the agent under study, and those who had reduced blood loss had lower operative mortality than those who did not, all other factors being equal, this also would support mechanism. It is important to note that unless the target step itself can be randomized, which is often not the case, this evidence will be derived from observational designs. If the intervention were randomly applied to one group and not another, and the clinical outcome observed, such direct empirical evidence would not constitute mechanistic evidence per se, although the demonstration that a difference in outcome correlated with the (nonrandom) difference in blood loss, the mechanism would be supported. Conversely, if blood loss in the two groups was identical, this would diminish the plausibility of conclusions derived from the direct empirical evidence, and precipitate a search for alternative mechanisms of clinical benefit.

A major question in nonhuman models is how to measure a clinical endpoint. By clinical endpoint is meant an outcome in a whole organism that in itself is likely to have a health benefit if experienced in humans. Such measures can include mortality, functional impairment, wasting and behavioral changes. Outcomes such as shrinkage of a tumor xenograft are more controversial, as such an effect in a human might be regarded as a surrogate outcome. This would depend on the specifics of the disease and therapy, and perhaps on the magnitude of the effect (e.g., complete remission versus modest reductions).

There is a varying language that is used in the literature for this step. In much of the *biomarker* and translational medicine literature establishing a clinical correlate for a change in the biomarker is called “qualification,” a step often preceded by “analytic validation” of the biomarker. (Altar, 2008; Alymani et al., 2010) However, in the surrogate endpoint literature this is frequently referred to as “clinical validation,” or simply “validation.” (Wagner, 2002; Kluft, 2004; Buyse et al., 2010) We will employ the latter usage but note that this refers to the same phenomenon as biomarker “qualification.” This dimension does not include the relevance of the outcome in the nonhuman model to a similar outcome in humans; that is captured in the next dimension.

As in the previous dimension, embedded within these evidentiary requirements are a set of components that are intervention or target independent, including components of good study design, and research on related interventions and targets. It includes both the qualitative considerations and the quantitative strength of relationships. Conventional measures of study quality apply; blinding of outcome assessment, use of multiple models in multiple laboratories, consistency of multiple related endpoints. Because many of these designs will be observational, standard criteria (e.g., Hill’s) for inferring a causal relationship from observational evidence can apply. If it is possible to induce the target effect through randomization, or if this can be achieved through mendelian randomization, the evidence is correspondingly stronger. (Altar, 2008; Alymani et al., 2010; Ransohoff, 2007, 2009)

Predictive Power of Nonhuman Model for an Effect in Humans

This dimension concerns the generalizability of evidence pertaining to an intervention’s mechanism of action. For this dimension to provide evidential weight there must be some basis for the extrapolation from animal to human based on mechanistic considerations. This is the “translation” step, and can be divided into the translation of the target effect (3t) and clinical effect (3c). Either could be based on a mechanistic basis for the analogy (i.e., all drugs with similar mechanism of action showed a relationship between animal and human effects). Such an analogy represents a chain of reasoning going back to the mechanism of action in at least one member of the class.

One of the concerns about the validity of animal to human extrapolation concerns whether the disease state in animals is produced by the same pathophysiologic processes as those of humans. These differences often arise from a poor understanding of these processes in humans, or of an inability to create or observe a parallel condition in nonhuman models. If the mechanisms of disease in humans are poorly understood it is very difficult to know whether the relevant components of that mechanism are adequately represented in nonhuman models. In addition, it can be extraordinarily difficult to mimic in nonhuman models those ancillary factors associated with human disease states that often mediate the effect of therapy, such as comorbidities, tolerance of treatment, and cotreatment.

In the end, the relevance of nonhuman (usually animal) models to humans is among the most difficult steps in the framework to satisfy or even to assess. The failure of animal models to reliably predict effects in humans has been widely documented (Alonso de Lecinana et al., 2001; 't Hart et al., 2004; DiBernardo and Cudkowicz, 2006; Benatar, 2007; Bolton, 2007; Crossley et al., 2008; Ayhan et al., 2009; Bracken, 2009). Appendix A, the annotated bibliography of literature on animal experimentation, documents the extent and causes of this issue. As noted previously, the failure to reliably predict does not mean that animal experiments do not provide substantial evidence; it merely means that they do not raise the probability high enough upon which to make clinical decisions. This predictive power varies widely by condition and treatment, but as noted previously, typically runs in the 5 percent to 20 percent range. Factors that have been described to diminish this predictive value include:

- Animal model for disease does not mimic human pathophysiology
- Design deficiencies in animal studies (e.g., related to randomization, allocation concealment and blinded outcome assessment, replication in multiple labs; Macleod et al., 2005)
- Uniformity of animal models not reflecting human phenotypic or genotypic variability.
- Different PK and PD properties in animals, improper dosing
- Differences in side effects, toxicity and drug interactions.
- Endpoints not mirroring human clinical endpoints
- Nonvalidated outcome measures

Target Effect in Human Disease States

The measure of the target effect is a biomarker. Therefore all of the factors that affect biomarker qualification, short of correlation with disease outcomes, contribute to evidence for this dimension. Evidence that the intervention can act at target site (e.g., absorption, PK/PD for drugs) is a precondition for creating a target effect. The evidence that the intervention is actually affect the target therefore requires:

- Validity of a measure of the target.
- Validity of measure of presence of intervention at target.
- Validity of measure of biologic effect on target.
- Evidence that intervention has desired impact on target effect measure.

The validity of the measures of each component of this step is a precondition to being able to assess whether a given intervention can alter the target state. When the intervention and target are readily observable, this is can be a trivial determination. However, for many devices and drugs, each of the above steps must be confirmed. For example, in determining whether an implantable cardiac defibrillator (ICD) has a biologic effect, the intervention is an electrical pulse in the myocardium, delivered by the ICD. The “target” is a ventricular arrhythmia, and the desired target effect is the conversion to normal ventricular rhythm. In the ICD case, the measure of intervention is the ICD firing record, the target measure is the intracardiac electrogram stored by the device, which also records the target effect, i.e., the conversion to normal ventricular rhythm. Both of these measures are validated. In this example, near instantaneous conversion is strong evidence of causality, although the proportion that would have continued to the point of physiologic compromise in the absence of the shock would still be uncertain. If there is an observed effect of ICDs on mortality without the devices firing, or without a commensurate number of converted arrhythmias, it becomes difficult to explain how the mortality effect was produced, weakening the evidence for the effect. In situations where causality is not evident in the individual case, the causal relationship must be ascertained in a randomized experiment with the target effect as the outcome. The ICD example also provides a mechanism for an adverse offsetting adverse effect, in that firing of the device in the absence of a target arrhythmia can increase mortality risk, akin to administering a drug where no pharmacologic target is present (Daubert et al., 2008).

In the drug context, the ability for the drug to affect the target step is best ascertained through a randomized or controlled experiment, in which the target effect is measured as a function of the active manipulation of the drug exposure. A detailed example of problems both with measurement and of assessing the target effect is provided in the subsequent Gleevec case study.

Clinical Impact of Target Effect in Human Disease State

The desiderata of this step in humans is essentially identical to that in nonhuman models, the main difference being that demonstration that the target effect is achievable in humans represent stronger evidence for the clinical effectiveness than does similar demonstration in nonhuman models. This step is equivalent to the biomarker "qualification" or surrogate endpoint "validation," as noted previously. It is often not possible to use randomization to completely eliminate the potential effect of covariates in determination of this effect. Target effects, which occur in some patients but not others, are typically post randomization events, and as such can require special statistical techniques to derive proper causal inferences (Prentice, 1989; Buyse et al., 2010).

Even with proper experimental design and analysis, the issue of the proper target measure to use to predict clinical effect can loom large. This was seen in the example of epidermal growth factor receptor (EGFR), where, as Taube et al. noted, "Detectable EGFR … appears to correlate with clinical benefit from EGFR inhibitors in some cases but fails to provide predictive information in others. It is unclear whether these differences are due to test methodologies, the biology of the disease being evaluated, or a combination of both." (Taube et al., 2009) So while the biologic measure of the target may be valid for the purposes of the precursor step, i.e., documenting the target effect, it may not be valid for the purposes of predicting therapeutic benefit. This underscores the criticality of recognizing that the validation of surrogate markers, measured in a specified way, as predictors of clinical impact often cannot be generalized beyond a given therapy, and clinical disease. When one strays beyond this, the evidence supporting this step has less weight.

A common error in the assessment of biomarkers is to fail to distinguish between the meaning of prognostic and predictive targets (Oldenhuis et al., 2008). A prognostic effect is one in which it is shown that a variation in the target, biomarker or surrogate endpoint correlates with a different frequency of a given clinical outcome. It should be stressed that this variation is not in itself a target "effect", which is rather an active change in the biomarker through intervention. For the assessment of therapeutic effects, the strength of evidence being sought is between this change, induced by the intervention, with clinical outcomes (Fleming and DeMets, 1996).

Finally, attention must be paid to whether the human model being used for this step is in fact one that is an accurate representation of the clinical setting in which the intervention will be applied. Is the target effect being assessed in patients with the disease or in healthy subjects? If in patients, do they have the same spectrum of comorbidities and cotreatments as expected in the clinical setting? Is the severity of disease similar? Is the mode and intensity (or dose) of administration of the intervention similar to the manner in which it will be applied clinically?

Step 5—Development of Case Studies and Pilot Test of the Framework

Two in depth case studies were developed to see how the conceptual framework being developed would apply to actual examples. The two case studies were of Gleevec for the treatment of chronic myelogenous leukemia and estrogen use in menopausal women for the prevention of heart disease. These case studies were summarized and mapped into the conceptual framework. These two examples are presented below.

Case Study 1: Gleevec (Imatinib)

General Background

We provide first a summary of the knowledge underlying the development of imatinib (Gleevec) for the treatment of chronic myeloid leukemia. Following this, we apply the framework to the facts of that case as they were known when the first human trials were begun.

Disease State

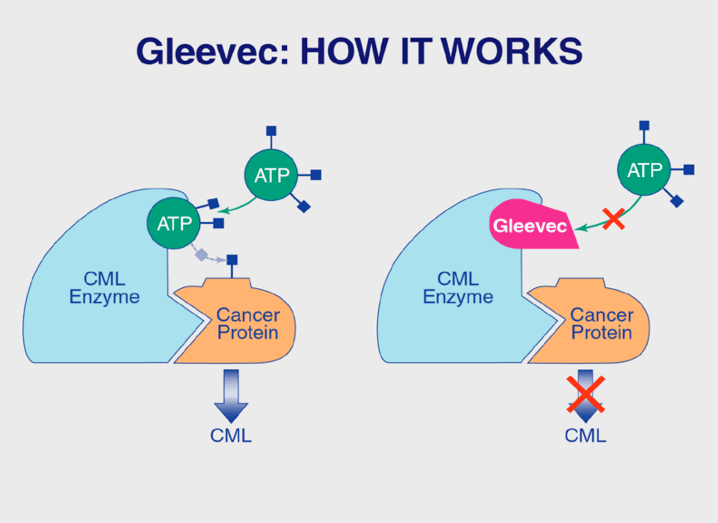
Chronic myeloid leukemia (CML) is a disease in which white blood cells are overproduced by the bone marrow. Normally, bone marrow cells called “blasts” mature into several different types of blood cells that have specific roles. CML affects the blasts that develop into white blood cells (granulocytes). The white blood cells do not mature normally, resulting in the proliferation of immature cells and interference with blood and marrow function. Most people with CML have an abnormal chromosome, known as the Philadelphia (Ph) chromosome, in which segments of chromosomes 9 and 22 are fused together. The molecular consequence of this event is the creation of a gene which codes for an abnormal protein called “Bcr-Abl.” This protein functions as a class of enzyme called a “tyrosine kinase,” producing cellular changes that result in CML.

In the majority of CML patients, there are no known hereditary, familial, geographic, ethnic, or economic associations with CML; therefore, the disease is not preventable nor does it appear to be heritable (Alvarez et al., 2007).

Therapeutic Mechanisms and Evidence

Gleevec or imatinib (also known as STI571) functions as a specific inhibitor of the Bcr-Abl tyrosine kinase (TK). The active sites of TKs have a binding site for adenosine triphosphate (ATP), which when bound triggers intracellular changes. The principal effect of Gleevec is to block Bcr-Abl’s ATP binding site, thereby inhibiting its enzymatic activity and ability to transform the cell. See Figure 2. Gleevec also induces apoptosis (cell death) of Bcr-Abl–expressing CML cells lines, halting the leukemic process. The target is therefore the Bcr-Abl binding site, and desired effect is to halt the cellular processes leading to carcinogenic transformation that would normally proceed if this enzyme is triggered.

Figure 2. Schematic model for the mechanism of action of Gleevec (imatinib)



Early preclinical and/or observational studies that lead researchers to explore Gleevec as possible intervention. Gleevec was developed from an observation first made by Nowell and Hungerford in 1960 that CML patients often had an unusually small chromosome (Philadelphia chromosome) not found in other people. Scientists had just learned how to isolate and study human chromosomes using newly developed cytogenetic techniques (the microscopic examination of genetic components of the cell, including chromosomes, genes, and gene products, using slides and stains).

At that point, the field theory of tumor development (i.e., the theory that neoplasms arose from many cells in a tissue made susceptible by exposure to carcinogenic agents) was still widely believed. It was also a time when most investigators did not think that tumors were caused by genetic mutations. Per Nowell (Nowell, 2007, p. 2035): “This might, in part, have been a reflection of the hope, more emotional than scientific, that tumors did not arise from structural changes in the genome because if this was the case, they would not be easy to reverse and treat.”

It was not until the 1970s that improving cytogenetic techniques and other methods were finally developed that generated more specific evidence at the level of individual chromosomes (implicating chromosomes 9 and 22 in CML), and ultimately led to the development of molecular techniques that permitted the identification of the specific genes altered not only in CML but in many hematopoietic and solid tumors.

Additional insight into the pathogenesis of CML came from studies (in mice) of transforming retroviruses. The transforming protein of this virus, v-Abl, was shown to be a tyrosine kinase (TK). The Bcr-Abl protein proved to display similar TK activity, which activated a variety of intracellular signaling pathways, leading to alterations in the proliferative, adhesive and survival properties of CML cells. In subsequent studies, the transfer of Bcr-Abl into mouse stem cells, followed by transplantation into mice, caused a CML-like syndrome; mice transgenic for Bcr-Abl developed acute leukemia. By 1990, these findings provided convincing evidence that Bcr-Abl was produced by a leukemic oncogene. A specific inhibitor of the Bcr-Abl protein was therefore predicted to be an effective therapeutic agent for CML.

Investigators screened chemical libraries to find a drug that would inhibit that protein. An initial lead compound (of the 2-phenylaminopyrimidine class) was identified by random, high-throughput screening; that is, the in vitro testing of large compound libraries. This lead compound was then tested and modified by the introduction of methyl and benzamide groups to give it enhanced binding properties, ultimately resulting in Gleevec.

Gleevec was subsequently tested in a number of preclinical models: in vitro, in vivo (mouse models), and human ex-vivo (peripheral blood or bone marrow from patients with CML). These experiments demonstrated that Gleevec specifically inhibits the proliferation of cell lines containing Bcr-Abl, and selects for the growth of benign hematopoietic progenitors in colony-forming assays using progenitor cells from CML patients. Studies in mice also showed that Gleevec had in vivo activity against Bcr-Abl-expressing cells and that continuous exposure to Gleevec was necessary to eradicate the tumors, suggesting that this would be important for optimal antileukemic effects. Prior to clinical testing, pharmacokinetic studies in mice, rats and dogs were conducted. Gleevec demonstrated favorable oral bioavailability, and lack of significant toxicity in animal models. Gleevec also demonstrated high selectivity for Abl TKs—leaving unaffected the dozens of other tyrosine kinases that are essential for normal function.

In 1998, the Phase I clinical trials of Gleevec were launched. Once doses of 300 mg or greater were reached, 53 out of 54 patients (98 percent) achieved a complete hematologic response (white blood cell, platelet, and red blood cell counts have returned to a normal range). At that dose level, cytogenetic responses (reduction in the number of cells in the bone marrow that have the Ph chromosome) were seen in 53 percent of patients, with 13 percent achieving a complete cytogenetic response. An attractive feature of CML for clinical concept validation is that the disease can be easily monitored through analysis of blood count, and the presence of the Ph chromosome can be used as a surrogate of disease activity. In this regard, the cytogenetic responses suggested that Bcr-Abl–expressing cells either had a selective growth disadvantage or they underwent cell death in the presence of Gleevec. These results have confirmed the role of Bcr-Abl and the importance of tyrosine kinase activity in the pathogenesis of CML.

How Thoroughly Characterized Was the Pathway?

In many respects, the disease pathway and the mechanism of drug action were thoroughly understood and described. However, several features of CML may make the success of a single agent such as Gleevec unique for this cancer. The Bcr-Abl tyrosine kinase, present in 95 percent of patients, is sufficient to cause the disease, and in early disease, it may represent the sole molecular abnormality. Few other malignant diseases can be ascribed to a single molecular defect in a protein kinase. In that regard it was an ideal target.

Even with this rather thorough understanding, there were and are still some unknowns. For example, at the time of Gleevec's development, no clear explanation for its impressive specificity for Abl-TK could be made. However, a series of recent biochemical and structural studies have elucidated the mechanisms responsible for the inhibition of the Bcr-Abl TK by Gleevec. TKs generally adopt similar active conformations, but can differ significantly in their inactive conformations; Gleevec inhibits Bcr-Abl specifically by binding to an *inactive* kinase domain conformation that is characteristic of Abl.

Gleevec displays excellent efficacy and minimal side effects with CML patients, and now represents the frontline therapy for CML. However, patients in advanced stages of the disease develop resistance to Gleevec treatment due to the acquisition of mutations in the Abl kinase domain that render the protein insensitive to this inhibitor. Second-generation drugs such as nilotinib and dasatinib have been developed that are able to target most, but not all, Gleevec-resistant mutations. Third-generation therapeutic agents are in development or clinical evaluation.

Finally, recent studies have revealed a potential "dark side" of Gleevec (Wang, 2006). Gleevec inhibits the normal cellular Abl, a downstream effector of the Eph receptors in breast cancer cells. These studies suggest that Eph-dependent tumor suppression requires Abl and is blocked by Gleevec, potentiating breast cancer promotion. Gleevec may not be overtly oncogenic, but its potential to promote some forms of tumor progression highlights the complexity of interactions between biological mechanisms that makes prediction of clinical effects from biological mechanisms difficult.

Application of Conceptual Framework to Gleevec

- **Condition**: Philadelphia chromosome positive Chronic myelogenous leukemia
- **Intervention**: Imatinib
- **Primary Target Effect**: Inhibition of abnormal tyrosine kinase activity
- **Clinical Effect**: Disease remission and cure

Strength of Evidence for Gleevec's Primary Target Effect in Nonhuman Models

Part of the background knowledge on leukemia was that it was a cellular, not systemic disease, i.e., if the abnormally proliferating cancer cells and their progenitors could be suppressed or eliminated, the disease would be controlled or cured. Imatinib emerged from the rational drug discovery process as the lead compound for preclinical development on the basis of its selectivity against CML cells in vitro and its drug-like attributes, including pharmacokinetic and formulation properties. Imatinib was the first drug in its class of targeted tyrosine kinase inhibitors.

Strength of Evidence for Gleevec's Clinical Effect in Nonhuman Models

Dose-dependent inhibition of tumor growth was seen in animals injected with human Bcr-Abl cells and treated daily with imatinib. Using a once-per-day schedule of up to 50 mg/kg, tumor growth was inhibited, but not eradicated (Druker, 1996). The reason for this modest in vivo activity became apparent from the pharmacokinetic profile of imatinib. This profiling revealed a short drug half-life in mice, which was not seen in other species (rat, dog, human). In nude mice a single dose of imatinib inhibited Bcr-Abl kinase activity for only 2 to 5 hours. A three times-per-day dosing schedule led to a continual block of Bcr-Abl kinase activity, resulting in eradication of tumors in 87 percent of imatinib-treated mice (le Coutre, 1999). On the basis of these data, it was considered likely that continuous exposure to imatinib would be required for optimal anti-leukemic effects.

Predictive Power of Nonhuman Model for an Effect in Humans

Several biological model systems demonstrated that Bcr-Abl is an oncogene that promotes CML pathogenesis. These model systems were important tools for elucidating the molecular mechanisms of CML formation, and to identify potential therapeutic targets. These included cultured cell models which demonstrated that the expression of Bcr-Abl could transform certain mouse cell lines and primary bone marrow cells (Ren, 2002). Animal models of in vivo CML pathogenesis demonstrated that the expression of Bcr-Abl in mouse bone marrow cells by retroviral transduction and bone marrow transplantation methods induced a myeloproliferative disorder that closely resembled CML (Daley, 1990; Kelliher, 1990; Elefanty, 1990). These models of CML proved that the Bcr-Abl kinase was sufficient to cause the disease, establishing it as the fundamental drug target.

Strength of Evidence for Target Effect in Humans

In a critical set of preclinical experiments conducted by Druker (Druker, 1996), imatinib was shown to suppress the proliferation of Bcr-Abl–expressing cells in vitro and in vivo. In colony-forming assays of peripheral blood or bone marrow from patients with CML, imatinib caused a 92 to 98 percent decrease in the number of Bcr-Abl colonies formed, with minimal inhibition of normal colony formation.

The standard dose-escalation, phase I study of imatinib involved escalation from 25 to 1000 mg in 14 cohorts of patients (Druker, 2001). Imatinib was rapidly absorbed after oral administration, and a mean maximal concentration was reached at steady state by once-daily administration of 400 mg. The half-life of the drug in the circulation ranged from 13 to 16 hours, and the levels of the drug increased by a factor of 2 or 3 at steady state with once-daily dosing. Blood samples from participants were tested to determine whether Bcr-Abl kinase activity was inhibited. The chosen measure was the mobility of CKRL, a major substrate of the Bcr-Abl enzyme. CRKL that is phosphorylated by BCR-ABL migrates more slowly on electrophoresis than the unphosphorylated form. Low doses (25 to 50 mg) of imatinib caused no alteration in the mobility of CRKL. An increase in the levels of the rapidly migrating unphosphorylated form and a concomitant decrease in the levels of the slowly migrating phosphorylated form were seen in patients receiving the 85-mg dose of imatinib; these changes were more prominent in patients receiving a daily dose of 140 mg and appeared to reach a plateau in patients receiving a daily dose of 250 to 750 mg.

Strength of Evidence for the Clinical Impact of the Target Effect in Human Disease States

In Phase I clinical trials, doses of 300 mg or greater of imatinib achieved a complete hematologic response (white blood cell, platelet, and red blood cell counts returned to a normal range) in 53 out of 54 patients (98 percent; Druker, 2001). Phase II trials were similarly successful (95 percent of patients achieved a complete hematologic response, 60 percent a major cytogenetic response, 13 percent relapse at median followup of 29 months), providing the basis for FDA approval (Kantarjian, 2002). In a 5-year followup study, 98 percent of patients showed a complete hematologic response, and the estimated overall survival rate for patients was 89 percent, with a relapse rate of about 17 percent (Druker, 2006).

How Well Does Conceptual Framework Capture Gleevec?

Imatinib is an ideal case to which to apply the framework. Unlike most other cancers, which are caused by a multitude of complex interacting genetic and environmental factors and therefore have many targets, CML is caused by a single aberrant protein related to a consistent chromosomal translocation. As such, researchers were able to focus all of their efforts on this single target. Nonetheless, the imatinib story is an excellent example of how evidence concerning biological mechanisms can lead to effective life-saving interventions. On balance, our conceptual framework does a good job in capturing the dimensions that contribute to the preclinical evidence in both predicting the successful clinical effects of Gleevec.

Case Study 2: Estrogen in Post-Menopausal Women

- **Intervention**: Hormone Replacement Therapy (HRT)
- **Primary Target Effect**: Lipoprotein Metabolism
    - Secondary Target Effects: Blood pressure, coagulation, and carbohydrate metabolism
- **Clinical Effect**: [Reduction in the risk of] Coronary Heart Disease (CHD)

Introduction

In the 1950s, evidence began to emerge that estrogen replacement therapy among postmenopausal women lowered the risk of coronary heart disease (CHD). Lipoprotein metabolism was considered the primary target effect of estrogen, and epidemiological, clinical, and laboratory studies supported this hypothesis. Still, there remained concerns about the validity of the observational data and the incomplete picture of the relevant biological mechanism(s). In the 1980s, the evidential picture became more complex with the addition of progestin to the estrogen regime in order to reduce the [increased] risk of endometrial cancer that was seen among estrogen users. As part of the Women's Health Initiative, which began in the 1990s, a randomized controlled trial (RCT) examined the effect of HRT on the risk of CHD.

Strength of Evidence for the Intervention's Primary Target Effect in Nonhuman Models

Much of the research on animal models investigating the relationship between estrogen, lipoprotein metabolism, and CHD took place in the 1950s and relied on rudimentary measures, such as total cholesterol. Estrogen was shown to reduce the total cholesterol:phospholipids (C/P) ratio in rats and cholesterol-fed male chicks, but not rabbits (Pick, Stamler, Rodhard, et al., 1952a, 1952b; Stamler, Pick, and Katz, 1956).

Strength of Evidence for Clinical Effect of Target Effect in Nonhuman Models

Consistent with the findings from research on estrogen's effect on lipoprotein metabolism, estrogen was shown to inhibit coronary atherogenesis and reverse previously induced coronary lesions among cholesterol-fed male chicks (Pick, Stamler, Rodhard, et al. 1952a; 1952b). Similarly, estrogens also resulted in lower levels of atherogenesis in rats but not rabbits (Stamler, Pick, and Katz, 1956).

Later research showed that estrogens reduced arterial lesions in female rabbits without significantly influencing plasma cholesterol levels, perhaps through direct interaction with the arterial wall or effects on plasma components other than lipoproteins (Hough and Zilversmit 1986). Although estrogen reduced the risk of atherosclerosis in sheep and cynomolgus monkeys, it did not impact lipoprotein levels, indicating that estrogen might operate through other mechanistic pathways to reduce the risk of CHD (Karas 2002).

Predictive Power of Nonhuman Model for an Effect in Humans

The use of gendered animal models also produced results consistent with early findings from clinical and epidemiological research. Egg-laying hens (modeling premenopausal women) had lower levels of atherogenesis and more favorable C/P ratios than male roosters, whereas ovariacteomized hens (modeling postmenopausal women) had levels of atherogenesis and C/P ratios that were similar to male roosters (Stamler, Pick, and Katz, 1956). However, for most mammalian models, estrogen appeared to exert any cardioprotective benefit through pathways *other* than its effect on the lipoprotein metabolism (Hough and Zilversmit 1986; Sarrel 1989).

The main issue with the use of animal models was the need to induce the disease state (i.e., hypercholesterolemia), usually through cholesterol feeding, in order to study the effects of estrogen on lipoproteins and CHD. In herbivores, such as rabbits, [total] cholesterol levels induced after cholesterol feeding were so high as to lead some to dismiss the model as irrelevant to the human disease state (Steinberg 2004). Other species, such as dogs or rats, were efficient at converting cholesterol and did not develop arterial lesions; this led some to argue against lipoproteins as a causative factor in CHD (Steinberg 2004).

Moreover, by the time progestins were added to estrogen replacement therapy, the importance of animal models in research on the relationship between HRT and CHD had declined. Evidence from animal models was weak and inconclusive regarding the effect of progestin on the risk of CHD via lipoprotein profiles.

Strength of Evidence for Target Effect in Humans (HRT, Lipid Protein Profile)

Clinical research produced conflicting results. In the 1970s, research had shown that while postmenopausal women had increased plasma levels of all lipoprotein patterns compared to premenopausal women, this pattern was similar to that of healthy men (Shoemaker, Forney, and MacDonald, 1977). Estrogen replacement did not result in a conversion to a premenopausal lipid profile. Rather, estrogen treatment was associated with a rise in HDL levels (potentially beneficial), a rise in VLDL levels (potentially harmful), and variable changes in LDL levels (Furman, 1971; Shoemaker, Forney, and MacDonald, 1977). Further research soon suggested that estrogen did in fact decrease LDL levels (Bush and Barrett-Connor, 1985).

However, research indicated that estrogen’s effect on lipids and lipid protein metabolism accounted for less than half of its purported cardioprotective effect (U.S. Food and Drug Administration, 1990; Espeland, et al., 1995). Other potential pathways included blood pressure, coagulation, and carbohydrate metabolism.

Progestins were shown to raise LDL levels and lower HDL levels, both of which could potentially increase the risk of CHD (Bush and Barrett-Connor, 1985; Bush, 1986; Watkins, 2007).

Strength of Evidence for the Clinical Impact of the Target Effect in Human Disease States

By the 1980s, the “lipid hypothesis”, namely the hypothesis that lipoprotein metabolism was a causative factor in the risk of CHD, was largely accepted by the scientific community. The results from the Lipid Research Clinics’ Primary Prevention Trial showed that reducing blood cholesterol with a pharmaceutical agent (cholestyramine) reduced the risk of CHD primary events (Steinberg, 2006). In 1984, an NIH consensus conference recommended that in order to reduce the risk of CHD, reducing blood cholesterol should be adopted as a national public health goal (Steinberg, 2006).

How Well Does the Framework Capture This Case Study?

The HRT-CHD case highlights important strengths and weaknesses of the conceptual framework. The case for the HRT-CHD hypothesis was constructed based on observational studies and research on surrogate endpoints, but there was insufficient evidence to determine whether HRT reduced the risk of CHD. The conceptual framework helps to systematically identify gaps in proposed mechanistic pathways and the types of evidence, mechanistic or otherwise, that should be generated to determine the effects of HRT on the risk of CHD. Indeed, if more attention had been given to the conflicting or inconclusive evidence concerning the effect of HRT on CHD via lipoprotein metabolism, the strong enthusiasm for the use of HRT to reduce the risk of CHD may have been tempered; as epidemiological and clinical evidence on surrogate endpoints mounted, less attention was given to understanding the basic mechanisms. On the other hand, the evidence suggested that HRT influenced the risk of CHD through multiple pathways (lipoprotein metabolism, blood pressure, and coagulation), and the conceptual framework does not provide guidance on how to integrate evaluations of the evidence across multiple mechanisms.

Discussion

We utilized multiple resources and perspectives including literature review and consultation with experts at our institution to develop a framework for the use of mechanistic knowledge in the evaluation of the effectiveness of medical interventions. This framework has several features combining work from a variety of fields that represent an important step forward in the rigorous assessment of such evidence.

- It uses a definition of evidence based on inferential effect, not study design.
- It separates evidence based on mechanistic knowledge from that based on direct evidence linking the intervention to a given clinical outcome.
- It represents the minimum sufficient set of steps for building an indirect chain of mechanistic evidence.
- It is completely adaptable and generalizable to all forms of interventions and health outcomes.
- It mirrors in the evidential framework the conceptual framework for translational medicine, thus linking the fields of basic science, evidence-based medicine and comparative effectiveness research.

Limitations and Future Research

While we believe the framework provided to be the starting point for any discussions of the value of mechanistic knowledge, much remains to be done in the form of both further refinement and implementation. In terms of refinement, while the framework components themselves represent a minimally sufficient set of dimensions, the optimal set of component questions within each of these dimensions requires further work. The more specificity that is provided in the sub questions, the more operational the framework becomes, but also potentially the more limited.

More work must also be done on how best to quantitate or weigh the impact both within and between various dimensions. Because many of the inferences cannot fall back on randomization, the same kinds of evidential judgments used when assessing observational studies must be applied to many of these designs. Building such a quantitative network or chain of inferences is akin to building complex risk models, and the relevance of such techniques to this application should be explored. As shown in some of the examples provided, it is possible to roughly quantitate the evidential value of the entire drug development process; refining this for specific interventions or in nondrug applications, will require substantially more work, yet is clearly achievable.

Most of the ideas presented herein are already part of the conversation of translational medicine and device and drug development. However, putting these together in the form of an evidential framework and a statistically sound definition of evidence is conceptually new territory for most working on either side of the translational divide. Developing a clear expert consensus, as there has been for traditional hierarchies of evidence, will require substantially more work. In addition, producing more clarity in how different forms of evidence interact will require further interdisciplinary research in both foundations and application. Bodies of work that can be mined for such development include that of expert elicitation and multiple bias modeling.

The pilot examples of the use of the framework demonstrated both its potential strength and areas for further work. It was clear in both cases the framework could be applied qualitatively, and that such application could illuminate those domains in which the evidence made the relationship between the therapy and the outcome more or less likely. In both cases, we saw that a limited number of pathways, a well characterized pathophysiology, and a clearly delineated target within those pathways were key elements. However, how the various qualitative observations can be quantitatively assessed and the relative weights of various dimensions or algorithmic combination thereof, requires further work.

It is important to note that judgments and decisions similar to those required of this framework are made every day in the assessment of newly evolving technologies. Whether a drug should proceed to clinical testing, whether a particular patient subgroup is more or less likely to respond to a drug, or whether weak direct evidence is sufficient in the face of strong mechanistic evidence to make therapeutic decisions, and finally, whether RCT evidence applies to a particular individual are judgments made on the basis of linkages and generalizations that have their grounding in mechanistic reasoning. Medicine has long known that mechanistic reasoning has its limitations in predicting the behavior of complex systems, but on the other hand, it has also shown that few therapies could have been developed or applied to individuals without such reasoning. It is not possible for medicine to reject such reasoning as a formal source of medical evidence, challenging as it is to formally assess and quantitate such information. We see the framework provided herein is the beginning of that process.

Conclusions

The formal language and logic of evidential assessment in evidence-based medicine and comparative effectiveness research has no formal place for incorporating knowledge of “how things work” in medicine. This project is provided a conceptual framework for that assessment, with proposals for how this might be combined with impure: direct evidence to provide a way of capturing all the ways of knowing in medicine, defined both on the group level and at the level of the individual. Much work remains to be done in terms of refining the subcomponents of these dimensions and in their categorization. Developing such a framework can help not only in the accurate representation of evidence for therapeutic decision-making and medical policy, but can potentially speed the development of medical interventions by demonstrating how and where mechanistic evidence can augment direct evidence. A potentially even more important outcome is that this framework can help bring together those communities working on the development and the assessment of therapies, who rarely seem to communicate except occasionally at the translational divide, and whose different views of what constitutes legitimate evidence has been a source of both misunderstanding and indeed conflict between those communities of researchers. Developing a common framework for evidence may be a first step towards true interdisciplinary, translational knowledge.

References

1. Alonso de Lecinana M, Diez-Tejedor E, Carceller F,et al. Cerebral ischemia: from animal studies to clinical practice. Should the methods be reviewed? Cerebrovasc Dis. 2001;11 Suppl 1:20-30.

2. Altar CA. The Biomarkers Consortium: on the critical path of drug discovery. Clin Pharmacol Ther. 2008 Feb;83:361-4.

3. Altar CA, Amakye D, Bounos D, et al. A prototypical process for creating evidentiary standards for biomarkers and diagnostics. Clin Pharmacol Ther. 2008 Feb;83:368-71.

4. Alvarez RH, Kantarjian H, Cortes JE. The biology of chronic myelogenous leukemia: implications for imatinib therapy. Semin Hematol. 2007 Jan;44(1 Suppl 1):S4-14. Review.

5. Alymani NA, Smith MD, Williams DJ, et al. Predictive biomarkers for personalised anti-cancer drug use: Discovery to clinical implementation. Eur J Cancer. 2010 Mar;46(5):869-79.

6. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004 Jun 19;328:1490.

7. Ayhan Y, Sawa A, Ross CA, et al. Animal models of gene-environment interactions in schizophrenia. Behav Brain Res. 2009 Dec 7;204:274-81.

8. Benatar M. Lost in translation: treatment trials in the SOD1 mouse and in human ALS. Neurobiol Dis. 2007 Apr;26:1-13.

9. Bolton C. The translation of drug efficacy from in vivo models to human disease with special reference to experimental autoimmune encephalomyelitis and multiple sclerosis. Inflammopharmacology. 2007 Oct;15:183-7.

10. Bracken MB. Why animal studies are often poor predictors of human reactions to exposure. J R Soc Med. 2009 Mar;102:120-2.

11. Bush TL. The adverse effects of hormonal therapy. Cardiol Clin. 1986 Feb;4(1):145-52.

12. Bush TL, Barrett-Connor E. Noncontraceptive estrogen use and cardiovascular disease. Epidemiologic Reviews. Epidemiol Rev. 1985;7:89-104. Review.

13. Buyse M, Sargent DJ, Grothey A, et al. Biomarkers and surrogate end points—the challenge of statistical validation. Nat Rev Clin Oncol. 2010 Jun;7(6):309-17.

14. Contopoulos-Ioannidis DG, Ntzani E, Ioannidis JP. Translation of highly promising basic science research into clinical applications. Am J Med. 2003 Apr 15;114:477-84.

15. Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. Stroke. 2008 Mar;39:929-34.

16. Daley GQ, Van Etten RA, Baltimore D. 1990. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. Science. 1990 Feb 16;247(4944):824-30.

17. Daubert JP, Zareba W, Cannom DS, et al. Inappropriate implantable cardioverter-defibrillator shocks in MADIT II: frequency, mechanisms, predictors, and survival impact. J Am Coll Cardiol. 2008 Apr 8;51:1357-65.

18. DiBernardo AB, Cudkowicz ME. Translating preclinical insights into effective human trials in ALS. Biochim Biophys Acta. 2006 Nov-Dec;1762:1139-49.

19. Druker BJ, Guilhot F, O'Brien SG, et al. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. N Engl J Med. 2006 Dec 7;355(23):2408-17.

20. Druker, BJ, Talpaz M, Resta DJ, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med. 2001 Apr 5;344(14):1031-7.

21. Druker BJ, Tamura S, Buchdunger E, et al. Effects of a selective inhibitor of the ABL tyrosine kinase on the growth of BCR-ABL positive cells. Nat Med. 1996 May;2(5):561-6.

22. Elefanty AG, Hariharan IK, Cory S. bcr–abl, the hallmark of chronic myeloid leukaemia in man, induces multiple haemopoietic neoplasms in mice. EMBO J. 1990 Apr;9(4):1069-78.

23. Espeland MA, Bush TL, Mebane-Sims I, et al. 1995. Rationale, design, and conduct of the PEPI trial. Control Clin Trials. 1995 Aug;16(4 Suppl):3S-19S.

24. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996 Oct 1;125:605-13.

25. Furman RH. Coronary heart disease and the menopause. In: Ryan KJ, Gibson DC. eds. Menopause and Aging. Washington, DC: U.S. Government Printing Office; 1971: 39-55.

26. Good IJ. Probability and the Weighing of Evidence. New York: Charles Griffin & Co.; 1950.

27. Goodman SN. Towards evidence-based medical statistics, II: The Bayes Factor. Ann Intern Med. 1999;130:1005-13.

28. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20:21-35.

29. Hill AB. The environment and disease: association or causation? Proc Roy Soc Med. 1965;58:295-300.

30. Hough JL, Zilversmit DB. Effect of 17 beta estradiol on aortic cholesterol content and metabolism in cholesterol-fed rabbits. Arteriosclerosis. 1986 Jan-Feb;6(1):57-63.

31. Kantarjian H, Sawyers C, Hochhaus A, et al. Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. N Engl J Med. 2002 Feb 28;346(9):645-52.

32. Karas RH. Animal models of the cardiovascular effects of exogenous hormones. Am J Cardiol. 2002 Jul 3;90(1A):22F-25F.

33. Kass RE, Raftery AE. Bayes Factors. JASA. 1995;90:773-95.

34. Kelliher MA, McLaughlin J, Witte ON, et al., Induction of a chronic myelogenous leukemia-like syndrome in mice with v-abl and BCR/ABL. Proc Natl Acad Sci U S A. 1990 Sep;87(17):6649-53.

35. Kluft C. Principles of use of surrogate markers and endpoints. Maturitas. 2004 Apr 15;47:293-8.

36. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2004 Aug;3:711-5.

37. le Coutre P, Mologni L, Cleris L, et al. In vivo eradication of human BCR/ABL-positive leukemia cells with an ABL kinase inhibitor. J Natl Cancer Inst. 1999 20;91:163-8.

38. Macleod MR, Ebrahim S, Roberts I. Surveying the literature from animal experiments: systematic review and meta-analysis are important contributions. BMJ. 2005 Jul 9;331:110.

39. Nowell PC. Discovery of the Philadelphia chromosome: a personal perspective. J Clin Invest. 2007 Aug;117:2033-5.

40. Oldenhuis CN, Oosting SF, Gietema JA, et al. Prognostic versus predictive value of biomarkers in oncology. Eur J Cancer. 2008 May;44:946-53.

41. Pick R, Stamler J, Rodhard S, et al. Estrogen-induced regression of atherosclerosis in cholesterol-fed chicks. Circulation. 1952a Dec;6(6):858-61.

42. Pick R, Stamler J, Rodhard S, et al. The inhibition of coronary atherosclerosis by estrogen in cholesterol-fed chicks. Circulation. 1952b Aug;6(2):276-80.

43. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. 1989 Apr;8:431-40.

44. Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. J Clin Epidemiol. 2007 Dec;60:1205-19.

45. Ransohoff DF. Promises and limitations of biomarkers. Recent Results Cancer Res. 2009;181:55-9.

46. Ren R. The molecular mechanism of chronic myelogenous leukemia and its therapeutic implications: studies in a murine model. Oncogene. 2002;21: 8629–42.

47. Royall R. Statistical Evidence: A Likelihood Paradigm. London: Chapman and Hall; 1997.

48. Shoemaker ES, Forney JP, MacDonald PC. Estrogen treatment of postmenopausal women: benefits and risks. JAMA 1977;238(14):1524-30.

49. Stamler J, Pick R, Katz LN. Experiences in assessing estrogen antiatherogenesis in the chick, rabbit, and man. Ann N Y Acad Sci; 1956: 596-619.

50. Steinberg D. Thematic review series: the pathogenesis of atherosclerosis.
An interpretive history of the cholesterol controversy, part V: the discovery of the statins and the end of the controversy. J Lipid Res. 2006 Jul;47(7):1339-51

51. Steinberg D. Thematic review series: the pathogenesis of atherosclerosis.
An interpretive history of the cholesterol controversy: part I. J Lipid Res. 2004 Sep;45(9):1583-93. Epub 2004 Apr 21.

52. Suppes P. A Probabilistic Theory of Causality. Amsterdam: Amsterdam; 1970.

53. Susser M. Judgment and causal inference in epidemiologic studies. Am J Epidemiol. 1977;105:1-15.

54. 't Hart BA, Amor S, Jonker M. Evaluating the validity of animal models for research into therapies for immune-based disorders. Drug Discov Today. 2004 Jun 15;9:517-24.

55. Taube SE, Clark GM, Dancey JE, et al. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. J Natl Cancer Inst. 2009 Nov 4;101:1453-63.

56. U.S. Food and Drug Administration. Fertility and Maternal Health Drugs Advisory Committee. Washington, DC: Federal Drug Administration; 1990: Meeting Minutes, 14 June 1990.

57. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet. 2004 May 22;363:1728-31.

58. Wagner JA. Overview of biomarkers and surrogate endpoints in drug development. Dis Markers. 2002;18:41-6.

59. Wang JY. Eph tumor suppression: the dark side of Gleevec. 2006 Aug;8(8):785-6.
Watkins ES. The Estrogen Elixir. Baltimore, MD: The Johns Hopkins University Press; 2007.

Appendix A. Annotated Bibliography of Animal Models Literature With Framework Mapping

This component of the project was conducted before the framework described in this document was fully developed and finalized. A preliminary framework was used into which the various articles were mapped. This mapping was done in the form of bolded codes that appear at the end of each article description, and correspond to the following dimensions.

1) **Strength of evidence for existence of intervention's pathway**
   a) Quality (design and execution) and strength (quantitative effect) of experimental evidence in preclinical models.
   b) Number of experimental models
   c) Variety of experimental models (e.g. animal species)
2) **Strength of evidence that the pathway exists in human disease states.**
   a) Strength of evidence for animal/in vitro model's relevance for human disease state.
   b) Ex vivo evidence
   c) Evidence that pathway occurs in complete physiologic system (e.g. functioning hearts vs. heart tissue.)
   d) Evidence from human physiologic experiments.
3) **Completeness of proposed mechanistic pathway.** (From intervention to clinical endpoint)
   a) Gaps in pathway (including whether intervention/exposure can exert effect on target due to issues of bioavailability, metabolism, delivery, etc.)
   b) Remoteness of the mechanistic outcomes from clinical outcomes.
   c) Strength of evidence linking proximal (i.e. surrogate) to distal (i.e. definitive) clinical endpoints
4) **Evidence for alternate, competing or compensatory pathways that can:**
   a) Produce outcome through pathways independent of intervention's effect
   b) Produce non-therapeutic outcomes through pathways dependent on intervention
   c) Interfere with intervention's pathways
5) **Strength of evidence that mechanism is similar to other interventions with known clinical effects**
6) **Adverse event mechanisms**

I. Annotated Articles

1. 't Hart BA, Amor S, Jonker M. Evaluating the validity of animal models for research into therapies for immune-based disorders. Drug Discov Today. 2004;9:517-24.

This article examines monoclonal antibody trials for immunotherapy in transplantation and for chronic diseases (rheumatoid arthritis and MS), and assesses the validity and predictive strength of animal models currently used for the development of effective therapies. The vast majority of immunology drugs have been preclinically tested in rodents, and, given the immunological differences between these mice and humans, it is not surprising that many fail to prove efficacious in humans. The authors argue that the outbred nature and immunological proximity of nonhuman primates to humans offer unique disease models to test whether the therapeutic principle holds in a higher species. **[1c, 2a, 3b]**

2. Alonso de Lecinana M, Diez-Tejedor E, Carceller F, Roda JM. Cerebral ischemia: from animal studies to clinical practice. Should the methods be reviewed? Cerebrovasc Dis. 2001;11 Suppl 1:20-30.

This article examines a number of preclinical focal cerebral ischemia models and discusses the reasons why findings from this research often fail to translate into clinically effective strategies. They present a number of explanations, including: the homogeneity obtained in animal models versus the high level of variability demonstrated among humans for critical pathological parameters of the condition; different PK properties; inattention to side effects and drug interactions in animal models; and methodological discrepancies, such as use of female, young animals and the use of endpoints that do not mirror clinical endpoints. The authors argue that these discrepancies must not invalidate preclinical studies. Rather, the knowledge of these reasons can help to optimize experimental models so that they become comparable with the clinical situation. **[1b, 2a, 3a]**

3. Anderson LM. Environmental genotoxicants/carcinogens and childhood cancer: bridgeable gaps in scientific knowledge. Mutat Res. 2006;608:136-56.

This article explores why, in numerous epidemiological studies, associations between childhood cancers and exposure to genotoxicants, including tobacco smoke, have been weak and hard to reproduce. The authors describe numerous scientific knowledge gaps and argue that conventional animal models should have a place in developing mechanistic understanding in filling these gaps. Perinatal bioassays in animals of specific environmental candidates, for example, benzene, could help guide epidemiology. Genetically engineered animal models could be useful for identification of chemical effects on specific genes. **[1b, 2a, 3a, 3b]**

4. Ayhan Y, Sawa A, Ross CA, et al. Animal models of gene-environment interactions in schizophrenia. Behav Brain Res. 2009;204:274-81.
5. Bailey GP, Marien D. What have we learned from pre-clinical juvenile toxicity studies? Reprod Toxicol. 2009;28:226-9.

This article assesses the scientific value of preclinical juvenile toxicity studies that are conducted to better predict the safety of pediatric drugs. The authors reviewed data from 10 pharmaceutical companies covering 39 studies. The authors found that only in 20 percent of the studies was it felt that the pre-clinical work contributed to the pediatric clinical trials and the preclinical studies were considered to have contributed to the product label in approximately 30 percent of cases. The authors raise questions about the need for clear scientific rationales in conducting these studies, suggesting that recently-implemented regulatory policies may be encouraging unnecessary and/or uninformative studies. **[1c, 6]**

6. Baker DH. Animal models in nutrition research. J Nutr. 2008;138:391-6.

This article reviews how experimental animal studies have contributed basic nutritional information concerning bioavailability of nutrients and nutrient precursors. It describes advantages, disadvantages and idiosyncrasies of numerous models, but does not offer a critical examination of them. **[1c, 3a]**

7. Bath PM, Macleod MR, Green AR. Emulating multicentre clinical stroke trials: a new paradigm for studying novel interventions in experimental models of stroke. Int J Stroke. 2009;4:471-9.

Building on the meta-analyses of neuroprotective agents in stroke led by Macleod, the authors argue for a fundamental paradigm shift away from performing preclinical studies in individual laboratories to performing them in an organized group of independent laboratories run by a steering committee and supported by a coordinating center, external data monitoring committee and outcome adjudication committee. This structure mimics the practice of multicenter RCTs. [**1a, 2a**]

8. Belser JA, Szretter KJ, Katz JM, et al. Use of animal models to understand the pandemic potential of highly pathogenic avian influenza viruses. Adv Virus Res. 2009;73:55-97.

This article reviews the advances made toward understanding the molecular determinants of avian influenza viruses. The use of mouse and ferret models has provided new insights into the contribution of virus and host responses and transmissibility, and in identifying the role of individual viral gene products and mapping the molecular determinants that influence the severity of disease. The article discusses the suitability of various animal models for their ability to reproduce human symptoms and pathogenesis (see Figure 1 in article). Authors argue that understanding the mechanisms of virulence of avian influenza viruses is crucial not only to develop improved antivirals and vaccines but also as a means to estimate the likely severity of disease for a given pandemic strain. **[1c, 2a]**

9. Benatar M. Lost in translation: treatment trials in the SOD1 mouse and in human ALS. Neurobiol Dis. 2007;26:1-13.

This article reports a meta-analysis of ALS treatment trials in mouse models and explores possible reasons for failure to translate promising preclinical findings into effective human treatments. While examining a number of reasons related to the methodological quality of these animal studies, the author also considers the relevance of the mouse model to human ALS, suggesting that the genetic mutation and time of treatment initiation used in most experiments are not relevant for the type of ALS (sporadic vs. familial) that the results are used to advance to human trials. **[1a, 2a, 3a]**

10. Bergman KL. The animal rule and emerging infections: the role of clinical pharmacology in determining an effective dose. Clin Pharmacol Ther. 2009;86:328-31.

This article examines drug development for emerging infections in translational pharmacology. Given the nature of emerging and re-emerging infections, specifically their severity (often life-threatening), the low incidence of natural occurrence even in endemic areas, and the potential of the infective agent to develop altered virulence and resistance to drugs, traditional drug development pathways (discovery, preclinical development, clinical development, post-approval) may not be possible. The Animal Rule — which allows the FDA to grant marketing approval based solely on animal studies if those studies are seen as providing substantial evidence of effectiveness in humans—is particularly relevant for the purposes of this report. Criteria for use of the Rule include: reasonably well-understood pathophysiological mechanism of toxicity; effects demonstrated in more than one animal species; end point clearly related to the desired benefit in humans, data or information on the pharmacokinetic/pharmacodynamics (PK/PD) of the product or other relevant data or information, in animals and humans, to allow selection of an effective dose in humans. **[1c, 2a, 3a, 6]**

11. Bodewes R, Rimmelzwaan GF, Osterhaus AD. Animal models for the preclinical evaluation of candidate influenza vaccines. Expert Rev Vaccines. 2010;9:59-72.

This article covers similar terrain as Belser et al (2009), number 8 in this appendix. Table 3 in the article, compares the advantages and disadvantages of various animal models most commonly used in the evaluation of candidate vaccines. It is informative in terms of the important predictive attributes of animal models viz. translation. **[1c, 2a, 3a]**

12. Bolton C. The translation of drug efficacy from in vivo models to human disease with special reference to experimental autoimmune encephalomyelitis and multiple sclerosis. Inflammopharmacology. 2007;15:183-7.

Similar to Friese et al. (2006, number 29 in this appendix), this article assesses preclinical models of experimental allergic (autoimmune) encephalomyelitis (EAE) and multiple sclerosis (MS), and provides some guidance that may improve clinical translation. The authors advocate for EAE models with representative and reproducible features, a uniform scoring system of disease, the inclusion of adequate controls, and careful choice of vehicle and an appreciation of the dose, route and frequency of treatment. They contend that the development of an accepted set of characteristics would provide a true picture of disease progression that could be used to confirm compound efficacy and ultimately help to counteract the discrepancies in drug activity between models and the corresponding human disease. **[1a]**

13. Bonjour JP, Ammann P, Rizzoli R. Importance of preclinical studies in the development of drugs for treatment of osteoporosis: a review related to the 1998 WHO guidelines. Osteoporos Int. 1999;9:379-93.

This article provides an overview of the World Health Organization osteoporosis guidelines, which underline the importance of a preclinical/clinical complementary program to assess the efficacy of new antiosteoporotic drugs. Preclinical studies carried out in the most reliable animal models (i.e., the most predictive with respect to human calcium and bone metabolism and drug responsiveness) are aimed at testing drug efficacy on bone mass/mineral density, microarchitecture and mechanical resistance in well-controlled conditions. The authors' review of animal studies indicated that these preclinical investigations were highly predictive of clinical outcome for most, if not all, drugs tested. The results of animal studies were able to predict whether changes in bone mass and/or bone mineral density were associated with modifications in bone fragility and therefore in fracture rate in osteoporotic patients. Preclinical studies also predicted the tolerance of bone tissue to increasing doses of the drugs, particularly with respect to the processes of modeling, remodeling, matrix mineralization and fracture healing. This is one of very few instances reporting success of preclinical animal models in terms of their ability to predict therapeutic efficacy. **[1a, 2a, 3c]**

14. Bracken MB. Why animal studies are often poor predictors of human reactions to exposure. J R Soc Med. 2009;102:120-2.
15. Bracken MB. Why are so many epidemiology associations inflated or wrong? Does poorly conducted animal research suggest implausible hypotheses? Ann Epidemiol. 2009;19:220-4.

In these two articles, Bracken suggests that the poor quality of animal research, and the way it is both synthesized and represented (dearth of systematic reviews; publication and outcome reporting biases), underlies the nonreplicability of many epidemiologic observations. **[1a, 3a]**

16. Chatzigeorgiou A, Halapas A, Kalafatakis K. The use of animal models in the study of diabetes mellitus. In Vivo. 2009;23:245-58.

This article provides a largely uncritical review and evaluation of rodent models of Types 1 and 2 diabetes. See Roep et al., 2004 for a more critical assessment of the limitations of rodent models, especially for Type 1 diabetes. **[1c]**

17. Corry DB, Irvin CG. Promise and pitfalls in animal-based asthma research: building a better mousetrap. Immunol Res. 2006;35:279-94.

The article reviews the challenges of animal models in asthma research. Given the complex disease process and heterogeneous pathogenesis of asthma, simple animal models have not reproduced in detail the underlying allergic immune mechanisms responsible for most forms of asthma and asthma-like diseases and correlate them with a limited set of clinically relevant disease variables. The authors suggest a number of technical improvements that could improve the reliability of experiments, but validity concerns (about disease initiation and exacerbation) persist and will only be resolved by continued animal experiments focused on understanding asthma pathophysiology. **[1c, 2a, 3a, 3b]**

18. Crossley NA, Sena E, Goehler J et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. Stroke. 2008;39:929-34.

The authors systematically identified and reanalyzed meta-analyses that described interventions in experimental stroke in order estimate the impact of various study quality items on efficacy estimates. They found that studies that failed to blind investigators and included healthy animals, as opposed to animals with comorbidities, overstated effect sizes. These findings are in keeping with this research group’s other results concerning study quality in the area of stroke. **[1a]**

19. Dehoux JP, Gianello P. The importance of large animal models in transplantation. Front Biosci. 2007;12:4864-80.

The authors review large animal models commonly used to evaluate organ transplant experiments and analyze the robustness of several models of human immune and physiological systems (especially allospecific tolerance and xenotransplantation). They suggest that rodent models be used to discover new genes and new biological pathways by using tools such as transgenic and knock-out animals. Large animal models should be used only to confirm findings; swine models seem to be the most appropriate choice, though nonhuman primate models may also provide relevant data. **[1c, 2a, 2c, 3a]**

20. DiBernardo AB, Cudkowicz ME. Translating preclinical insights into effective human trials in ALS. Biochim Biophys Acta. 2006;1762:1139-49.

This article provides an overview of important features in the discovery, development, and validation of disease-modifying therapies and interventions for ALS. The animal (especially mouse) models for ALS thus far have failed to predict response in humans. The reasons for discordant results between mouse and human trials may relate to inherent differences between the mouse and human disease (comparability of pharmacokinetics, routes of delivery, timing of treatment, relevance of familial disease model to sporadic disease). Despite these failures, authors assert that mouse models remain important tool in pursuing new therapeutic approaches. **[3a, 3b]**

21. Dirnagl U, Macleod MR. Stroke research at a road block: the streets from adversity should be paved with meta-analysis and good laboratory practice. Br J Pharmacol. 2009;157:1154-6.
22. Dirnagl U. Bench to bedside: the quest for quality in experimental stroke research. J Cereb Blood Flow Metab. 2006;26:1465-78.

While this article is largely focused on the poor quality of preclinical stroke research, the author discusses some critical translational hurdles, including: species differences, inappropriate time windows of treatment, effective drug levels not achievable in humans because of toxicity, use of young animals without comorbidity, failure to model white matter damage and protect axons, incongruent end points, and heterogeneity of stroke subtypes in patients, among others. The article (indirectly) raises questions about important trade-offs between reductionist mechanistic models that may be important for basic, narrow discoveries versus more complex models that may better mirror human disease state. **[1a, 3a, 3b]**

23. Dixon JA, Spinale FG. Large animal models of heart failure: a critical link in the translation of basic science to clinical practice. Circ Heart Fail. 2009;2:262-71.

This article provides an overview of a number of animal models and species used preclinical research on heart failure (including recent developments in gene therapy and stem cells), highlighting the utility and value of large animal models. The authors suggest that large animal models have often played a critical role in successful translation from bench to bedside. They caution that recent advances in our understanding heart failure at the molecular and protein levels will not result in successful translation without large animal models that recapitulate the clinical heart failure phenotype in ways that murine models cannot. **[1c, 3a]**

24. Dragunow M. The adult human brain in preclinical drug development. Nat Rev Drug Discov. 2008;7:659-66.

This article takes as its starting point the fact that no effective neuroprotective agent (save tissue plasminogen activator (TPA) has been developed for humans for neurodegenerative disorders such as Alzheimer’s disease and Parkinson’s disease. Authors suggest that animal models are necessary for neuroprotective drug development (especially dose selection and toxicological assessment) but are not sufficient. Animal models of human brain disorders by necessity tend to focus on and therefore model specific aspects of the disease, and cannot reproduce the complex array of human neuropathology and symptomatology. They recommend expanding target validation by using human brain tissue microarray screening and direct adult human brain cell testing at an early preclinical stage (an adult human brain preclinical platform) to isolate molecules that protect the human brain (see Figure 2 in article). **[2b, 4a]**

25. Dyson A, Singer M. Animal models of sepsis: why does preclinical efficacy fail to translate to the clinical setting? Crit Care Med. 2009;37:S30-7.

This article takes as a starting point that preclinical models of sepsis (largely utilizing mice and rodents) cannot replicate the complexity of human sepsis. Disparities in severity of insult, species, comorbidities, gender, and age make translation difficult. While the authors do not suggest alternatives to animal models, they argue that the models themselves are too heterogeneous, and recommend using standardized animal models as way of improving translation. **[1b, 2a, 3c]**

26. Ferrante RJ. Mouse models of Huntington's disease and methodological considerations for therapeutic trials. Biochim Biophys Acta. 2009;1792:506-20.

The author of this article reviews some of the successful developments in the use of genetic mouse models of Huntington's disease, and carefully considers what constitutes sufficient data from mouse models to justify translation to humans. Experiments with these models have yielded many promising therapeutic candidates, but there is a need to prioritize these leads. Given the variability of lab procedures and models, it can be exceedingly difficult to compare evidence of efficacy and effect size. The author offers numerous methodological recommendations that will allow for more rigorous selection of leads for human trials. **[1a]**

27. Fielden MR, Kolaja KL. The role of early in vivo toxicity testing in drug discovery toxicology. Expert Opin Drug Saf. 2008;7:107-10.

This opinion piece focuses on in vivo preclinical toxicity testing and suggests that the predictivity of these models is lacking. The authors recommend larger upfront investment in experiments designed narrowly to obtain a more thorough understanding of the mechanisms of toxicity, arguing that the current experimental paradigm—focused on efficacy and PK properties—does not sufficiently or meaningfully inform the selection and prioritization of compounds. They propose an alternative early testing strategy (Figure 1 in article) that will shift attrition of future failing molecules upstream in the discovery process. **[6]**

28. Fisher M, Henninger N. Translational research in stroke: taking advances in the pathophysiology and treatment of stroke from the experimental setting to clinical trials. Curr Neurol Neurosci Rep. 2007;7:35-41.

This article by one of the developers of the STAIR criteria summarizes some lessons learned from preclinical stroke research to date. Figures 1 (STAIR recommendations) and 2 (lessons learned) within the article's text provide a useful summary. **[1a, 3c]**

29. Friese MA, Montalban X, Willcox N, et al. The value of animal models for drug development in multiple sclerosis. Brain. 2006;129:1940-52.

The rodent model typically used in preclinical MS studies—induced EAE— does not reproduce all the pathogenetic mechanisms operating in spontaneous human MS. MS is highly heterogeneous in its genetic basis, environmental effects, clinical course, pathological mechanisms, and treatment responsiveness, and this heterogeneity needs to be comprehended and mimicked in any ideal animal model (Box 1 within article). The authors are hopeful that the use of more humanized mouse models (using transgenic and stem cell technologies) that incorporate multiple susceptibility factors may reproduce the clinical heterogeneity of MS better, and improve identification of promising therapeutic approaches. **[1a, 3a]**

30. Gallegos RP, Nockel PJ, Rivard AL, et al. The current state of in-vivo pre-clinical animal models for heart valve evaluation. J Heart Valve Dis. 2005;14:423-32.

This article provides an overview of current animal models of preclinical safety evaluation of prosthetic heart valves developed for use in humans. The authors endorse the use of standard sheep models, which in their estimation most accurately simulates most characteristics of human anatomy and physiology. Of particular relevance to BMEBM is the inclusion of the International Standards Organization guidance in formulating ideal animal studies, summarized in the article in Table 1. **[1c, 3b, 6]**

31. Ganter B, Giroux CN. Emerging applications of network and pathway analysis in drug discovery and development. Curr Opin Drug Discov Devel. 2008;11:86-94.

While not strictly about the predictivity of animal models, this article discusses recent applications of pathway and network analysis for predictive in silico modeling in the area of drug discovery and development. These tools link relevant extracted literature information (including reports of animal experiments) with features that enable analysis and interpretation of the global impact of a disease stage or drug treatment. Such integrated models can link cellular profiles of genomics, proteomic and metabolomic data with the corresponding clinical endpoint, and can provide a new perspective for drug discovery and development. Figure 1 in article describes the workflow embodied in this approach, and includes the role of in vivo preclinical data. **[1b, 2a, 2b, 4a]**

32. Geerts H. Of mice and men: bridging the translational disconnect in CNS drug discovery. CNS Drugs. 2009;23:915-926.

This paper reports on a number of under-appreciated fundamental differences between animal models and human patients in the context of drug discovery with emphasis on Alzheimer's disease and schizophrenia. These differences include the absence of many functional genotypes in animal models and difficulties in simulating the pre-ymptomatic state (Figure 1in article). The author offers possible solutions to these translational challenges, including organizational improvements (information and cost-sharing collaborations), the better use of negative trial data, technical improvements (development of better imaging biomarkers), the introduction of realistic drug schedules early in drug discovery, and the use of computational models (Figure 2 in article). At bottom, however, the biggest improvements in translation will result from new conceptual models that treat CNS disorders as imbalances of networks rather than mismatches of single targets, and multi-target molecules that may lead to significant clinical improvements. **[1a, 3a]**

33. Gold R, Linington C, Lassmann H. Understanding pathogenesis and therapy of multiple sclerosis via animal models: 70 years of merits and culprits in experimental autoimmune encephalomyelitis research. Brain. 2006;129:1953-71.

This article covers similar terrain as Friese et al. (2006, number 29 in this appendix) and Bolton (2007 number 12 in this appendix) in reviewing EAE models of MS. While acknowledging the limitations of many of the models, the authors argue that they have resulted in many advances of our understanding and treatment of MS, and represent the best hope for further progress in MS treatment. Table 1, which summarizes commonly used rodent models of EAE and their similarities to and differences from human disease, is particularly useful. **[1b, 3a, 3b]**

34. Gurwitz D, Weizman A. Animal models and human genome diversity: the pitfalls of inbred mice. Drug Discov Today. 2001;6:766-8.

This article summarizes the limitations of using inbred mice in preclinical experiments, especially the use of single strains that do not reflect the natural variation of the human patient population. The authors support the development of a mouse genome project that would eventually allow genome-wide comparative genomic studies, leading to the identification of new drug targets that share similar natural variations in mice and humans and, thus, are more suitable for studies in mouse models for human diseases. **[1b, 2a]**

35. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. *JAMA*. 2006;296:1731-2.
36. Hackam DG. Translating animal research into clinical benefit. BMJ. 2007;334:163-4.

The above two brief papers by Hackam emphasize the poor methodological quality of animal studies. In the systematic review of 76 highly cited animal studies, the authors found that only just over a third translated at the level of human randomized trials, a rate of translation is lower than the estimated 44 percent replication rate for highly cited human studies in Ioannidis (2005). Hackam recommends uniform reporting requirements and rigorous systematic reviews of animal experiments prior to human trials as potential solutions. **[1a]**

37. Hausheer FH, Kochat H, Parker AR, et al. New approaches to drug discovery and development: a mechanism-based approach to pharmaceutical research and its application to BNP7787, a novel chemoprotective agent. Cancer Chemother Pharmacol. 2003;52 Suppl 1:S3-15.

In the face of poor predictivity of animal models and the "serendipity" of the compound screening process, the authors recommend an alternative approach to drug discovery, based on the elucidation and exploitation of biological, pharmacological, and biochemical mechanisms that have not been previously recognized or fully understood. Mechanism-based drug discovery (MBDR) involves the combined application of physics-based computer simulations and laboratory experimentation. MBDR research is based on the following principle: if a series of molecular simulations of the properties of a biological target, chemical transformations, stability and interactions, or drug–target interactions of interest are in agreement with a series of experimental observations of the molecular systems of interest, the corresponding probability that such observations are true and correct is greatly increased. This approach is aimed at reducing the probability of failure and enhancing the development process. **[1b, 2b, 4a]**

38. Hein WR, Griebel PJ. A road less traveled: large animal models in immunological research. Nat Rev Immunol. 2003;3:79-84.

The authors argue that in immunological research there has been too much dependence on a single lab species (mice) that are, in critical ways, biologically irrelevant to the study of human disease and the development of therapies. They recommend placing greater emphasis on biological relevance and making use of large(r) animal models. **[1c]**

39. Herodin F, Thullier P, Garin D, et al. Nonhuman primates are relevant models for research in hematology, immunology and virology. Eur Cytokine Netw. 2005;16:104-16.

Like Hein and Griebel (2003), the authors argue that the great similarity of nonhuman primates (NHPs) to humans justifies their use in the investigation of pathophysiological mechanisms in hematology, immunology and virology and in the evaluation of tolerance and efficacy of candidate therapeutics. Rodents are not sufficiently relevant to be able to predict human responsiveness to biological modifiers, pathogens, and potential therapeutics, notwithstanding the advantages conferred by the diversity of transgenic and knock-out murine models. Following a screening step in rodents, the availability of sophisticated cell and gene therapy tools makes it compulsory to validate them in preclinical trials with NHPs. **[1c]**

40. Hersch SM, Ferrante RJ. Translating therapies for Huntington’s disease from genetic animal models to clinical trials. NeuroRx. 2004;1:298-306.

The article examines what constitutes an informative genetic animal model (in neurological disease generally, and Huntington’s disease in particular), what principals should be followed in designing experiments using genetic models, and what constitutes sufficient mechanistic evidence to justify translation to humans. It includes useful discussion about importance of distinguishing between primary outcomes (neuropathological evidence of neuroprotection) and secondary outcomes (related to symptoms of Huntington’s disease). The impact therapeutic trials in genetic models can have on selecting compounds for clinical trials in humans depends on many factors relating to the quality and breadth of the preclinical data, captured in Figure 1 of the paper. **[1a, 1c, 2a, 3b]**

41. Horrobin DF. Modern biomedical research: an internally self-consistent universe with little contact with medical reality? Nat Rev Drug Discov. 2003; 2(2):151-4.

The author suggests that that biomedical science, and hence pharmaceutical science, has taken a wrong turn in its relationship to human disease. The information generated by cell culture, animal models of disease, transgenic mice and molecular biology studies rests on faulty and frequently unexamined assumptions and is not congruent with the “real world of medical illness.” Animal models “represent nothing more than an extraordinary, and in most cases irrational, leap of faith.” If we are to continue using animal models, at the very least we ought to test our assumptions by constantly referring back to the original disease in humans. **[1a, 2a]**

42. Hsu CY. Criteria for valid preclinical trials using animal stroke models. Stroke. 1993 May; 24(5): 633-6.

This editorial addresses numerous challenges regarding study design and quality in the area of stroke research. The author suggests that the shortcomings inherent to clinical trials are often absent in animal experiments (lack of more objective outcome measures, diversity in stroke pathology, heterogeneity of demographic factors, comorbidities, variable delay in starting treatment). He argues that animal experiments should be held to the same rigorous design and conduct standards in place for clinical trials. **[1a]**

43. Insel TR. From animal models to model animals. Biol Psychiatry. 2007;62:1337-9.

This editorial makes two points. First, it argues that biological psychiatry can learn much from modern comparative neurobiology, which studies the neural basis of species-typical behaviors rather than looking for phenocopies of human behavior. Second, it argues that traditional animal models might be mechanistically misleading, but the experimental use of model organisms (chosen strategically to test hypotheses) to understand the pathophysiology of mental disorders will be critical as clinical studies identify genetic alleles and cellular changes that confer risk for mental disorders. **[1b, 2a]**

44. Jeffery EH, Keck AS. Translating knowledge generated by epidemiological and in vitro studies into dietary cancer prevention. Mol Nutr Food Res. 2008;52 Suppl 1:S7-17.

The article examines the lack of preclinical evidence in dietary cancer prevention, which lead to clinical trials that “provide confusing, disappointing, and maybe even harmful results.” The authors argue that once a poorly designed clinical trial fails to demonstrate a proposed benefit, it can take years and several trials to correct. They suggest that mechanistic evidence from in vitro studies and animal modeling of efficacy, bioavailability, and kinetics are essential for designing robust clinical trials. Figure 1 in the article describes authors’ view of standard and optimal approaches to scientific study of foods with health benefits. **[1a, 2a, 3a]**

45. Joers VL, Emborg ME. Preclinical assessment of stem cell therapies for neurological diseases. ILAR J. 2009;51:24-41.

This article reviews the requirements of stem cell-based therapy for clinical translation, advances in stem cell research toward clinical application for neurological disorders, and different animal models used for analysis of these potential therapies (focusing on Parkinson’s disease, stroke and MS). Of particular interest for BMEBM is the discussion of the challenges in demonstrating the efficacy and safety of grafting human stem cells in animal models. **[1c, 3a, 6]**

46. Kamat CD, Gadal S, Mhatre M, et al. Antioxidants in central nervous system diseases: preclinical promise and translational challenges. J Alzheimers Dis. 2008;15:473-93.

Recent high-profile failures of vitamin E trials in Parkinson’s disease, and nitrone therapies in stroke, have diminished enthusiasm to pursue antioxidant neuroprotectants in the clinic. The authors carefully consider whether the failures result from antioxidant theory or the implementation of that theory. The argue that evidence for the theory’s validity is convincing, but evidence of implementation flaws abound, including failure to understand the drug candidate’s mechanism of action in relationship to human disease, and failure to conduct preclinical studies using concentration and time parameters relevant to the clinical setting. **[1a, 2a, 3b]**

47. Kirschvink N, Reinhold P. Use of alternative animals as asthma models. Curr Drug Targets. 2008;9:470-84.

This review focuses on the availability, advantages and nonadvantages of asthma models in nonlaboratory animals (cats, dogs, sheep, swine, cattle, horses, and monkey). The authors advocate for the use of these large animals because they offer the great potential to perform long-term functional studies allowing a simultaneous within-subject approach of functional, inflammatory and morphological changes. **[1c]**

48. Knight A. Animal experiments scrutinised: systematic reviews demonstrate poor human clinical and toxicological utility. ALTEX. 2007;24:320-5.
49. Knight A. Systematic reviews of animal experiments demonstrate poor contributions toward human health care. Rev Recent Clin Trials. 2008;3:89-96.
50. Knight A. Reviewing existing knowledge prior to conducting animal studies. Altern Lab Anim. 2008 Dec; 36(6): 709-12.

In the above three papers, the author challenges the assumption that animal models provide an predictive basis which would justify their use in toxicity testing and biomedical research aimed at developing cures for human diseases. To investigate the validity of this assumption, he conducted a search of SCOPUS databases for published systematic reviews of the human clinical or toxicological utility of animal experiments. Of 20 reviews examining clinical utility, authors concluded that the animal models were substantially consistent with or useful in advancing clinical outcomes in only 2 cases. Possible causes include interspecies differences, the distortion of experimental outcomes arising from experimental environments and protocols, and the poor methodological quality of many animal experiments. While the latter problems might be minimized, the interspecies limitations may be technically and theoretically impossible to overcome. Yet, unlike nonanimal models, animal models are not normally subjected to formal scientific validation. The author argues that instead of simply assuming they are predictive of human outcomes, the consistent application of formal validation studies to all test models is clearly warranted. **[1a, 2a]**

51. Ledford H. Translational research: the full cycle. Nature. 2008;453:843-5.

This journalistic article examines the notion of reverse translation—that clinical trials and patients' unexpected responses are valuable human experiments, and failed trials can stimulate new hypotheses that may help refine the experiment in its next iteration. This "bedside to bench" approach is explained through the recounting of three clinical trials (cancer drug, gene therapy, HIV vaccine). **[1a]**

52. Lemon R, Dunnett SB. Surveying the literature from animal experiments. BMJ. 2005;330:977-8.

In this editorial, the authors take the view that a review of all known relevant preclinical experiments should be conducted prior to human clinical trials. They recommend performing what they call a "critical review" rather than a systematic review. A critical review compiles and evaluates the different sources of experimental evidence on a qualitative basis. A difficulty with systematic reviews is that attempts to meet precise inclusion criteria often mean useful information is excluded. They argue that the reliability and validity of each animal model needs to be assessed on its merits and its relevance to the particular clinical application. **[1a]**

53. Linder S, Shoshan MC. Is translational research compatible with preclinical publication strategies? Radiat Oncol. 2006;1:4.

In this paper, the authors examine translational difficulties in the area of cancer therapeutics. They argue that a number of factors contribute to making the translation process inefficient, including the use of sensitive cell lines and fast growing experimental tumors as targets for novel therapies, and the use of unrealistic drug concentrations and radiation doses. They suggest that the aggressive interpretation of data, successful in hypothesis-building biological research, does not form a solid base for development of clinically useful treatment modalities, and question whether "clean" results obtained in simplified models, expected for publication in high-impact journals, represent solid foundations for improved treatment of patients. They recommend increasing open-access publishing to increase dissemination and transparency of all relevant data. **[1a, 3a]**

54. Lindner MD. Clinical attrition due to biased preclinical assessments of potential efficacy. Pharmacol Ther. 2007;115:148-75.

This article examines the magnitude and prevalence of numerous biases that may affect preclinical assessments of potential efficacy. The author argues that the shift to more target-based drug discovery has increased bias, suggesting that proof of concept studies that used to be conducted fairly early, before strong attachments to individual targets had developed, are now conducted at the end of the lead optimization phase, 3 to 5 years into the program, at a point when considerable time and resources have already been invested. He recommends a number of ways to limit bias (cultural, procedural, decision-making). **[1a]**

55. Loscher W. Preclinical assessment of proconvulsant drug activity and its relevance for predicting adverse events in humans. Eur J Pharmacol. 2009;610:1-11.

This article compares preclinical and clinical models for the assessment of proconvulsant activity of investigational or marketed drugs. The author argues that a major limitation of tests to assess the safety of various agents is the specific mechanism of action of convulsant effect, so that testing of drugs may produce both false positive and false negative data, and argues for a different set of tests that can provide complete and more reliable conclusions about the proconvulsant potential of an investigational drug. These tests should include animals with lowered seizure threshold, and consider the relation of doses producing (pro)convulsant effects to the therapeutic dose-range of a substance (“therapeutic index”). **[1b, 6]**

56. Lowenstein PR, Castro MG. Uncertainty in the translation of preclinical experiments to clinical trials: Why do most Phase III clinical trials fail? Current Gene Therapy. 2009; 9: 368-74.

This paper assesses why so few Phase III clinical trials have failed in translation from preclinical experiments. It briefly describes some of the complications of preclinical experimentation generally (availability of numerous types of models, each with own advantages and disadvantages; human patients having been exposed to the “standard of care” prior to the novel therapy; statistical issues, including over-reliance on p<0.05 and failure to analyze effect size); genetic homogeneity of experimental animals; scaling; disease time course). The authors suggest that a main limitation of the basic science is the "lack of comprehensive understanding of which variables being examined are actually significant and/or rate limiting parameters that are relevant to the study of human disease, and predictive of novel treatments' efficacy in human patients." They provide recommendations for how the process from preclinical experiments to RCTs can be made more "robust," defined as an experimental system's ability to "maintain its central functions in the face of challenges." They recommend preclinical testing in a variety of models in different genetic backgrounds, ages, sizes, and species, to show whether efficiency seen in a homogenous genetic background is robust viz. genetic heterogeneity. They also recommend that early phase trials should be designed to simultaneously target safety and treatment efficacy, not just safety as is currently the case. The authors do not propose specific ways of better capturing/evaluating preclinical evidence, but suggest that developments in mathematical, statistical and biological models will allow for more rigorous assessment of such evidence. **[1a, 1b, 3a]**

57. Lynch VJ. Use with caution: developmental systems divergence and potential pitfalls of animal models. Yale J Biol Med. 2009;82: 53-66.

The author of this article challenges the assumption that gene functions and genetic systems are conserved between models and humans, arguing that evidence that gene functions and networks diverge during evolution is often overlooked. A number of mechanisms that generate functional divergence and recent examples demonstrating that gene functions and regulatory networks diverge through time are presented. The author argues that the examples suggest that annotation of gene functions based solely on mutant phenotypes in animal models, as well as assumptions of conserved functions between species, can be wrong. Therefore, animal models of gene function and human disease may not provide appropriate information, particularly for rapidly evolving genes and systems. **[2a]**

58. Macleod MR, Ebrahim S, Roberts I. Surveying the literature from animal experiments: systematic review and meta-analysis are important contributions. BMJ. 2005;331:110.
59. Macleod MR, Fisher M, O'Collins V et al. Good laboratory practice: preventing introduction of bias at the bench. Stroke. 2009;40:e50-2.
60. Macleod MR, O'Collins T, Howells DW, et al. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke. 2004;35:1203-8.
61. Macleod MR, van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke. 2008;39:2824-9.

In the above four papers, Macleod and colleagues present empirical evidence of bias and poor study quality in the area of ischemic stroke. On the contrary, Lemon and Dunnett (2005) argue that quantitatively-oriented systematic reviews and meta-analyses are preferable to a "critical" review approach. They propose a series of measures/practices aimed at reducing bias in preclinical stroke experiments including randomization, allocation concealment, sample size calculations, and blinded assessment of outcome. The two systematic reviews of neuroprotective agents demonstrate that preclinical reports of efficacy are confounded by study quality biases. **[1a]**

62. Malkesman O, Austin DR, Chen G, Manji HK. Reverse translational strategies for developing animal models of bipolar disorder. Dis Model Mech. 2009;2:238-45.

The article highlights a number of issues relevant to BMEBM. One is that the phenotypical complexity of human disease, particularly in the case of bipolar disorder (BD), is rarely captured in preclinical animal models, which rely on simpler phenotypes. The authors use three criteria—face validity, predictive validity, and construct validity—to evaluate animal models in BD, suggesting that construct validity allows researchers to generate a possible common mechanistic theory that can explain both the animal model and the human disorder. They suggest using construct validity, rather than face validity, as a starting point for creating models, capitalizing on technological advances that allow researchers to create animal models that reflect the biological changes observed in studies of individuals with BD. They believe that this strategy, while imperfect, will help to support valid hypotheses regarding the mechanisms of BD. **[1a, 1b, 3a]**

63. Manger PR, Cort J, Ebrahim N, et al. Is 21st century neuroscience too focused on the rat/mouse model of brain function and dysfunction*? Front Neuroanat*. 2008;2:5.

This paper presents an analysis that demonstrates that 75 percent of neuroscience research efforts are directed to the rat, mouse, and human brain, or 0.0001 percent of the nervous systems on the planet. This extreme bias in research trends may provide a limited scope in the discovery of novel aspects of brain structure and function that would be of importance in understanding both the evolution of the human brain and in selecting appropriate animal models for use in clinically relevant research of mental illnesses. **[1b, 2a]**

64. Manto M, Marmolino D. Animal models of human cerebellar ataxias: a cornerstone for the therapies of the twenty-first century. Cerebellum. 2009;8:137-54.
65. Manto M, Marmolino D. Cerebellar disorders—at the crossroad of molecular pathways and diagnosis. Cerebellum. 2009;8:417-22.

These two articles provide a largely uncritical review of developments concerning preclinical models of cerebellar ataxias. These models have yielded significant breakthroughs in our understanding of the pathogenesis of cerebellar ataxias (especially at molecular level), reproducing to various extents human brain disorders. The authors are hopeful that these findings will be integrated into clinical research and that therapeutic strategies will move beyond merely the treatment of symptoms. **[1c, 2a]**

66. Mao J. Translational pain research: achievements and challenges. J Pain. 2009;10:1001-11.

This article reviews the advances made in recent pain research and examines the translational gaps between pain mechanisms and clinical pain. The author considers potential causes of these gaps, both from bench to bedside (experimental conditions, PK/PD issues such as dosage and bioavailability, discrepancy in pain assessment tools, comorbidity/gender/genetic differences) and bedside to bench (experimental pain models, spontaneous vs. stimulus-induced pain, acute vs. chronic pain). The author identifies the development of objective pain-assessment tools as a fundamentally important goal of pain research. **[1a, 3a]**

67. Markou A, Chiamulera C, Geyer MA, et al. Removing obstacles in neuroscience drug discovery: the future path for animal models. Neuropsychopharmacology. 2009; 34:74-89.

The article discusses the traditional role of animal models in neuroscience drug discovery (focused mainly on psychiatric, as opposed to neurological disorders) and the reasons why this approach has led to suboptimal utilization of the information that animal models provide. Certain experiments and recombinant DNA technologies (creating knockout mice) are widely-used, but their predictive validity for clinical benefit has not been critically examined. Preclinical and clinical measures need to assess as closely as possible homologous, or at least analogous, biological variables. The authors argue that such correspondence between preclinical and clinical measures will greatly enhance predictability, and thus promote translation back and forth between animal and human studies. Furthermore, the measures used both preclinically and clinically should have construct validity, defined as measuring accurately the theoretical behavioral and neurobiological variables that are considered core to the disorder of interest. **[1a, 2a]**

68. Marshall JC, Deitch E, Moldawer LL, et al. Preclinical models of shock and sepsis: what can they tell us? Shock. 2005;24 Suppl 1:1-6.

The authors of this paper argue that while preclinical models of shock and sepsis do not predict therapeutic efficacy in human disease, they provide insights that may be of use in deciding whether a strategy is worth evaluating in the clinical arena, and if so, in which patients and under what circumstances. These models can also point to potential adverse effects that may limit the use of that strategy in particular groups of patients. Table 4 in the paper outlines an approach to the development of a portfolio of preclinical models that is especially insightful. **[1a, 3b, 6]**

69. Matthews RA. Medical progress depends on animal models—doesn't it? J R Soc Med. 2008;101:95-98.

The author proposes a calculation to assess the evidential weight provided by animal models. This can be done using the concepts of sensitivity (i.e., the true positive rate) and specificity (i.e., true negative rate). These lead to various ways of quantifying evidential weight, of which the most direct and transparent is the likelihood ratio (LR), whose definition is such that only tests producing LR >1 can be deemed to have contributed any weight of evidence. The paucity of quantitative comparative data for animal models makes even such simple calculations impossible. The author offers numerous explanations for this paucity: (1) compounds that produce unacceptable effects in animal models will not progress to human trials, making studies capable of giving sensitivity/false positive rates for animal models ethically problematic; (2) it is frequently difficult to establish end-points sufficiently clear-cut to allow categorization as true positives or true negatives; (3) much of the comparative animal-human data is obtained under conditions of commercial confidentiality. **[1a]**

70. Miczek KA, de Wit H. Challenges for translational psychopharmacology research—some basic principles. Psychopharmacology (Berl). 2008;199:291-301.

This thoughtful paper lays out a number of principles for translating preclinical findings to clinical applications in the area of psychopharmacological drug development. The key challenge – particularly acute in research on psychiatric disorders — is that few models of psychiatric disorders are homologous with the disorder; rather the laboratory procedures model isomorphic signs and symptoms. The principles of note for BMEBM purposes include:

1. The translation of preclinical data to clinical concerns is more successful when the scope of experimental models is restricted to a core symptom of a psychiatric disorder.
2. Preclinical experimental models gain in clinical relevance if they incorporate conditions that induce maladaptive behavioral or physiological changes that have some correspondence with species-normative behavioral adaptations.
3. Preclinical data are more readily translated to the clinical situation when they are based on converging evidence from several experimental procedures, each capturing cardinal features of the disorder.
4. The more closely a model approximates significant clinical symptoms, the more likely it is to generate data that will yield clinical benefits.
5. The choice of environmental, genetic, and/or physiological manipulations that induce a cardinal symptom or cluster of behavioral symptoms reveals the theoretical approach used to construct the model.
6. Preclinical experimental preparations that are validated by predicting treatment success with a prototypic agent are only able to detect alternative treatments that are based on the same mechanism as the existing treatment that was used to validate the screen.
7. The degree to which an experimental model fulfills the criteria of high construct validity relative to face or predictive validity depends on the purpose of the model. **[1b, 2a, 3b]**

71. Mignini LE, Khan KS. Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. BMC Med Res Methodol. 2006;6:10.

This paper reports a review of systematic reviews of animal studies and found the methodological rigor of the systematic reviews lacking in terms of their assessment of study validity and quality. The authors found that reviews often lacked methodological features such as specification of a testable hypothesis, assessment of publication bias, study validity and heterogeneity, and meta-analysis for quantitative synthesis. They assert that there is a need for more rigor in reviewing animal research. **[1a]**

72. Mitchell BF, Taggart MJ. Are animal models relevant to key aspects of human parturition? Am J Physiol Regul Integr Comp Physiol. 2009;297:R525-45.

This article critically reviews the data and concepts concerning the use of animal models for parturition and offers a rationale for the use of a new model. A number of animal models have contributed to advances in understanding the regulation of parturition. The authors suggest that animals dependent on progesterone withdrawal to initiate parturition clearly have a limitation to their translation to the human. In these models, a linear sequence of events gives rise to a "trigger" mechanism. The authors propose that human parturition arises from the maturation of several systems in parallel, and emphasize the need to determine the precise role of the immune system in the process of parturition. They support the development of nonprimate animal models whose physiology is more relevant to human parturition (guinea pig) and who display key physiological characteristics of gestation that more closely resemble human pregnancy than do currently favored animal models. **[1c, 2a]**

73. Mogil JS. Animal models of pain: progress and challenges. Nat Rev Neurosci. 2009;10:283-94.

This paper reviews the state of the art regarding behavioral animal models of pain. There is a useful discussion and defense of why animal models are needed in this area of research, which includes a brief discussion of clinical face validity. Box 1 in the paper offers some conceptual clarification, re: what we mean in using the term “animal model,” distinguishing between the subject, the assay, and the measure. The authors also highlight the disconnect between preclinical experiments (where young, male animals are used) and the epidemiological evidence of human pain (typical chronic-pain patient is middle-aged and female). **[1a, 2a, 3a]**

74. Muschler GF, Raut VP, Patterson TE, et al. The design and use of animal models for translational research in bone tissue engineering and regenerative medicine. Tissue Eng Part B Rev. 2010 Feb;16(1):123-45.

This article provides an overview of animal models for the evaluation, comparison, and systematic optimization of tissue engineering and regenerative medicine strategies related to bone tissue. It includes an overview of major factors that influence the rational design and selection of an animal model. Two sections of the paper are of import to BMEBM. One describes “missing links” between preclinical and clinical performance, including: underestimation of variation in clinical response, overestimation of performance, and insensitivity to incremental improvement. In a section (re: “gaps and opportunities “ to improve existing models) the authors identify gaps in the availability of animal models, including: (1) the need for assessment of the predictive value of preclinical models for relative clinical efficacy, (2) the need for models that more effectively mimic the wound healing environment and mass transport conditions in the most challenging clinical settings, and (3) the need for models that allow better measurement and detection of cell trafficking events and ultimate cell fate. **[1c, 3a, 3b]**

75. O’Collins VE, Macleod MR, Donnan GA, et al. 1,026 experimental treatments in acute stroke. Ann Neurol. 2006;59:467-77.

This systematic review sought to identify agents tested in animal neuroprotection models and those treatments given to acute stroke patients; and to compare the overall quality of evidence and experimental efficacy of those treatments that have been given to acute stroke patients and those agents that have not progressed beyond the experimental phase. The numerous findings and recommendations related to poor study design and quality are significant. All told, there was no evidence that drugs used clinically were more effective experimentally than those tested only in animal models. Moreover, no particular mechanism of action in animal models demonstrated superior efficacy, leading the authors to suggest that the current stroke models are in need of reformulation. The authors argue that intervention should be considered for clinical trial only when there is both a high level of experimental efficacy and a diverse body of evidence supporting its clinical application. **[1a, 2a]**

76. Opal SM, Patrozou E. Translational research in the development of novel sepsis therapeutics: logical deductive reasoning or mission impossible? Crit Care Med. 2009;37:S10-5.

Like Marshall et al. (2005), the authors highlight some of the translational challenges of sepsis research. They discuss a number of technological advances that may allow for more realistic technology recapitulation of events in the pathophysiology of sepsis, which may assist in the preclinical evaluation of antisepsis drugs. In the short term, they advocate using the PIRO concept (predisposing factors, infection type, host response, and organ dysfunction model) to deal with the multiple parameters that affect outcome in sepsis (see Table 2 in paper). Animal models should take at least some of these factors in consideration in the design of preclinical programs to study new antisepsis agents. **[1b, 2a, 3a]**

77. Pacharinsak C, Beitz A. Animal models of cancer pain. Comp Med. 2008;58:220-33.

This article reviews a number of recently developed models of cancer pain. While earlier models examined anatomic mechanisms, recent models (mostly rodent, but some feline and canine models) are examining basic biochemical, molecular, and neurobiologic mechanisms. These models — which allow researchers to generate novel hypotheses regarding the roles of genes and their protein products in pain processing and modulation — will be crucial to developing novel therapeutic drugs that specifically target particular genes for specific types of cancer pain. **[1c, 2a, 2d, 3a]**

78. Palena C, Abrams SI, Schlom J, Hodge JW. Cancer vaccines: preclinical studies and novel strategies. Adv Cancer Res. 2006;95:115-45.

This article reviews findings from preclinical cancer vaccine studies conducted in animal tumor models. While progress in understanding the molecular mechanisms of immune activation has helped in the design of novel and more efficient vaccine strategies, the authors contend that major translational challenges remain. One is related to the relevance of the utilized models. Most preclinical work to date has been conducted with transplanted murine tumors that grow rapidly, are usually noninvasive, and fail to metastasize. Most human tumors grow slowly and do not represent the percent of body mass that murine tumors do. The short time span of mouse models precludes multiple booster vaccinations, so few cycles of vaccine immunotherapy can be given. This is in contrast to the vaccine therapy in a patient with minimal residual disease, who can receive many cycles of immunotherapy over the course of several years. A second challenge is related to the fact that many defined tumor antigens are self-proteins and therefore generally fail to initiate strong antitumor T-cell responses. Thus, a key for developing successful cancer vaccines is to overcome potential mechanisms of immune suppression against antigenic but weakly immunogenic tumors. **[1c, 2a, 3a]**

79. Pegram M, Ngo D. Application and potential limitations of animal models utilized in the development of trastuzumab (Herceptin): a case study. Adv Drug Deliv Rev. 2006;58:723-34.

This article presents a case study of trastuzumab with a focus on the role of animal models in many phases of the drug’s development. The authors review what was learned from murine models to understand the pathogenesis of breast cancer, test efficacy of various monoclonal anti-HER2 antibodies, and to provide insight into the mechanism of action of the drug. The principle shortcoming of animal modeling in the development of trastuzumab was the lack of cross reactivity of trastuzumab to nonhuman HER2, making it difficult, if not impossible, to predict unanticipated toxicities such as cardiac dysfunction. **[1a, 3a, 6]**

80. Perel P, Roberts I, Sena E, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ. 2007;334:197.

The authors conducted meta-analyses of all available animal data for six interventions that showed definitive proof of benefit or harm in humans. For three of the interventions—corticosteroids for brain injury, antifibrinolytics in hemorrhage, and tirilazad for acute ischemic stroke—they found major discordance between the results of the animal experiments and human trials. Equally concerning, they found consistent methodological flaws throughout the animal data, irrespective of the intervention or disease studied. In addition, the use of randomization, concealed allocation, and blinded outcome assessment—standards that are considered the norm when planning and reporting modern human clinical trials—were inconsistent in the animal studies. **[1a]**

81. Peters JL, Sutton AJ, Jones DR, et al. A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. J Environ Sci Health B. 2006;41:1245-1258.

This systematic review examines the extent and quality of systematic reviews and meta-analyses of in vivo animal experiments carried out to inform human health. They found a number of methodological and reporting deficiencies in both SRs and MAs, and propose a modified QUOROM or MOOSE guidelines specific to animal experiments. **[1a]**

82. Philip M, Benatar M, Fisher M, et al. Methodological quality of animal studies of neuroprotective agents currently in phase II/III acute ischemic stroke trials. Stroke. 2009;40:577-81.

In similar fashion to the studies by Macleod and colleagues, this paper analyzes the quality and adequacy of animal studies supporting the efficacy of NXY-059 and other neuroprotective agents investigated in phase II/III trials. The authors identified the reports of animal experiments in the Phase II/III studies and applied five STAIR criteria to evaluate the quality of each report. They also examined the collective literature for each individual drug to determine the range of experiments that were performed. Sufficiency of the preclinical literature for each drug was evaluated using a set of criteria derived from five other STAIR criteria. The authors found substantial within-drug and between-drug variability in the methodological quality of the published studies and insufficient preclinical data for all of the drugs in phase II/III trials. **[1a]**

83. Pienta KJ, Abate-Shen C, Agus DB, et al. The current state of preclinical prostate cancer animal models. Prostate. 2008;68:629-39.

This article identifies a number of discovery bottlenecks that have impeded the translation of preclinical prostate cancer animal models, including: (1) insufficient number of models with insufficient molecular and biologic diversity to reflect human cancer, (2) a lack of understanding of the molecular events that define tumorigenesis, and (3) failure to address why preclinical studies appear not to be predictive of human clinical trials. With regard to (3), the authors advocate for preclinical studies that utilize the appropriate agent doses, and pharmacokinetic and pharmacodynamic parameters to take into account the differences in metabolism between mouse and human. They argue for improved feedback in the design of both preclinical studies, which would include thinking about how the agents can be given in humans, and the design of clinical trials, which rarely take into account how the preclinical testing was accomplished. **[1a, 1c, 2a, 3a]**

84. Piper RD, Cook DJ, Bone RC, et al. Introducing Critical Appraisal to studies of animal models investigating novel therapies in sepsis. Crit Care Med. 1996;24:2059-70.

While the disease addressed here—sepsis — is covered more comprehensively in other articles (Dyson 2009; Marshall 2005, Opal 2009), this article outlines an evidence-based approach to the assessment of preclinical animal studies evaluating novel therapeutic interventions. The "levels-of-evidence" approach proposed in the paper (see Tables 2, 3, and 5 for criteria re: study assessment, study selection, and evaluation of the literature) is instructive for BMEBM purposes. **[1a]**

85. Pound P, Ebrahim S, Sandercock P, et al. Where is the evidence that animal research benefits humans? BMJ. 2004;328:514-7.

This article examines published systematic reviews of animal experiments, focusing on reviews that had been conducted to find out how the animal research had informed the clinical research. The authors found that the results of only one—thrombolytics for acute ischemic stroke—showed similar findings for humans and animals (and this was for similar excess risk for intracranial hemorrhage). They identify several methodological problems of animal experiments, including: disparate animal species and strains, different models for inducing illness or injury with varying similarity to the human condition, use of a variety of outcome measures, which may be disease surrogates or precursors and which are of uncertain relevance to the human clinical condition, absence of randomization and blinding. The authors argue that systematic reviews should become routine to ensure the best use of existing animal data as well as improve the estimates of effect from animal experiments. **[1a, 1c, 2a, 3c]**

86. Rice AS, Cimino-Brown D, Eisenach JC, et al. Animal models and the prediction of efficacy in clinical trials of analgesic drugs: a critical appraisal and call for uniform reporting standards. Pain. 2008;139:243-7.

This article complements Mogil (2009), proposing a number of refinements in the way animal experiments are conducted in pain research. These refinements include: matching the animal model to the disease that is proposed for the eventual clinical indication, squaring the outcomes of interest between animal studies and clinical trials, and designing the animal studies to more closely resemble the human experience of neuropathic pain (chronicity, incidence, comorbidity, late onset). **[1a, 2a, 3b]**

87. Ritter T, Nosov M, Griffin MD. Gene therapy in transplantation: Toward clinical trials. Curr Opin Mol Ther. 2009;11:504-12.

This article examines why despite many promising studies the translation of preclinical gene therapy strategies to clinical trials has been minimal. The authors contend that there has been reluctance among transplant researchers to initiate trials involving gene therapy, one reason being that the immunological and nonimmunological mechanisms underlying acute and chronic transplant failure are highly complex, and there is a perception that the manipulation of a single genetic target is unlikely to improve the outcomes of the majority of organ transplant recipients. Figure 1 presents a useful schematic representation of progress toward the clinical application of five different gene therapy approaches, from proof-of-principle to controlled clinical trials. **[2a, 3a]**

88. Roberts I, Kwan I, Evans P, et al. Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. BMJ. 2002;324:474-6.

This systematic review of animal experiments on fluid resuscitation found that most studies were underpowered and provided little information on possible bias. The authors call for improvements in the design and quality of animal experiments and for the increased use of rigorously conducted systematic reviews of animal experiments. **[1a]**

89. Roep BO. Are insights gained from NOD mice sufficient to guide clinical translation? Another inconvenient truth. Ann N Y Acad Sci 2007 Apr 1103:1-10.
90. Roep BO, Atkinson M, von Herrath M. Satisfaction (not) guaranteed: re-evaluating the use of animal models of type 1 diabetes. Nat Rev Immunol. 2004 Dec;4(12):989-97.

In the two articles above the authors argue that since rodent models of type 1 diabetes (T1D) have failed to determine the precise mechanisms of disease initiation/progression and to inform design of interventions that prevent or cure T1D, a philosophical change in preclinical research is needed. This reorientation would include studies in controlled environments, interventional analyses across a broad range of times and doses, robust studies of safety, considerations of genetic and immunological differences, studies carried out in more than one animal model, and frequent comparison with emerging human data. Table 3 provides a very useful "roadmap" to improved use of animal models. **[1a, 1c, 2a]**

91. Rosenblum WI. Criteria for valid preclinical trials using animal stroke models. Stroke. 1993;24:1601-2.

In this brief letter responding to Hsu (1993), the author makes the point that reproducibility/ repeatability of experiments is critical to establishing valid, robust animal models. He argues that the disinclination of many journals to publish confirmatory studies does a disservice to biomedical research enterprise. **[1a]**

92. Schnabel J. Neuroscience: Standard model. Nature. 2008;454:682-5.

This article provides a journalistic account of the problems concerning poor study quality and questionable validity of preclinical models of neurodegenerative diseases. **[1a, 2a]**

93. Schook L, Beattie C, Beever J, et al. Swine in biomedical research: creating the building blocks of animal models. Anim Biotechnol. 2005;16:183-90.

This conference report argues that the opportunities for utilizing swine biomedical models are immense, particularly in models that address lifestyle issues (nutrition, stress, alcohol, drugs of abuse, etc.). The authors suggest that in order to fully capitalize upon the promise, there needs to be greater recognition of cofactors, such as nutrition, as key modulators of phenotype via genomic, epigenetic, and postgenomic mechanisms. **[1c, 2a]**

94. Segalat L. Invertebrate animal models of diseases as screening tools in drug discovery. ACS Chem Biol. 2007;2:231-6.

The article examines trade-offs between highly relevant but high throughput screening (HTS)-incompatible mammalian models and poorly predictive in vitro models and explores the merits of invertebrate disease models. They authors recommend using invertebrate animals in HTS research contexts, acknowledging that such models are imperfect but useful in some research contexts, summarized in Table 1. The main advantage of invertebrates over other in vitro assays is that they provide a system that is both HTS-compatible and in which the physiological context is preserved. **[1b, 2a]**

95. Sena E, van der Worp HB, Howells D, et al. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci. 2007;30:433-9.

This article suggests that reports of the efficacy of candidate neuroprotective drugs in animal models of stroke are profoundly biased by aspects of study design. Conclusions drawn from individual publications or from narrative reviews cannot provide the basis for selecting drugs for clinical trial or for the design of those clinical trials. A sound judgment on efficacy, the limits to efficacy, the need for any further animal experiments and the design of any ensuing clinical trial can only be made on the basis of a systematic analyses of all available animal data; such an analysis must include the possible contribution of publication and study-quality bias to the observed efficacy. The authors evaluate checklists for evaluating both the range of evidence for efficacy and individual study quality (Table1). **[1a]**

96. Shanks N, Greek R, Greek J. Are animal models predictive for humans? Philos Ethics Humanit Med. 2009;4:2.

The authors of this philosophical article provide a conceptual analysis of the term "scientific prediction" and contend that there is no credible empirical evidence that animal models can predict human responses to drugs. They argue that reliance on faulty causal analogical reasoning and the conflation between sensitivity and specificity are two sources of error that lead researchers to have misplaced confidence in the predictive utility of animal models. They do not offer any alternatives, which renders the article a rather verbose exercise in hand-wringing. **[1a]**

97. Soubret A, Helmlinger G, Dumotier B, et al. Modeling and simulation of preclinical cardiac safety: towards an integrative framework. Drug Metab Pharmacokinet. 2009;24:76-90.

This review article describes essential components of cardiac electrophysiology modeling and simulation. The authors propose that a progressive integration of such mechanistic components into a common quantitative framework may help improve understanding and predictability of drug-induced TdP risk. Preclinical studies have provided a deeper understanding of torsadogenic mechanisms and potential pro-arrhythmic markers to assess. Translating preclinical insights into a quantitative clinical risk assessment remains challenging because of (i) species differences in cardiac electrophysiology and drug pharmacokinetics; and (ii) the inability to measure clinically specific cardiac electrophysiology metrics, and therefore ascertain the full predictive value of earlier preclinical components of the risk assessment process. **[1a, 3b, 6]**

98. Suzuki Y, Yeung AC, Ikeno F. The pre-clinical animal model in the translational research of interventional cardiology. JACC Cardiovasc Interv. 2009;2:373-83.

This review provides an overview of the emerging results of preclinical studies and development, and evaluation of animal models for percutaneous cardiovascular device technologies for patients with symptomatic cardiovascular disease. **[1c, 2a]**

99. Swanson KS, Mazur MJ, Vashisht K, et al. Genomics and clinical medicine: rationale for creating and effectively evaluating animal models. Exp Biol Med (Maywood). 2004;229:866-75.

The recent advent of techniques in molecular biology, genomics, transgenesis, and cloning furnishes investigators with the ability to study vertebrates (e.g., pigs, cows, chickens, dogs) with greater precision and utilize them as model organisms. Comparative and functional genomics and proteomics provide effective approaches for identifying the genetic and environmental factors responsible for complex diseases and in the development of prevention and treatment strategies and therapeutics. By identifying and studying homologous genes across species, researchers are able to accurately translate and apply experimental data from animal experiments to humans. This review supports the hypothesis that associated enabling technologies can be used to create, de novo, appropriate animal models that recapitulate the human clinical manifestation. Comparative and functional genomic and proteomic techniques can then be used to identify gene and protein functions and the interactions responsible for disease phenotypes, which aids in the development of prevention and treatment strategies. Figure 1 in the paper, which describes methods to choose, create, and interpret data generated from animal models is particularly useful for BMEBM purposes. **[1b, 1c, 2a, 3a]**

100. Thannickal VJ, Roman J. Challenges in translating preclinical studies to effective drug therapies in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2010 Mar: 181(6):532-3.

This editorial was written in response to a report of a randomized, placebo-controlled Phase II clinical trial of the safety and efficacy of imatinib for the treatment of idiopathic pulmonary fibrosis (IPF). The study failed to demonstrate a difference in the primary end-point of disease progression, leading the authors to examine the preclinical evidence that supported moving to clinical trials. In IPF, the accurate determination of the "dominant" aberrant cellular phenotype involved in disease pathogenesis and the associated/altered signaling pathway(s) is not sufficiently well defined, and the roles of several other cells and phenotypes also remain unclear.

The authors propose some minimal preclinical criteria to be applied before beginning clinical trials of novel IPF agents, including: (1) identification of the targeted molecule and the activation of its related signaling pathway in human IPF lung tissues when compared with appropriate controls; (2) demonstration that the candidate drug/therapeutic agent modulates the specific cellular profibrogenic phenotype(s) in animal models and in cells obtained from human IPF lungs; and (3) demonstration of the antifibrotic effects of the agent in at least two different animal models of lung fibrosis, with drug being delivered during the postinflammatory, fibrogenic phase of lung injury.
[1a, 1c, 2a, 2b, 3a]

101. Thrusfield MV. Ageing in animal populations—an epidemiological perspective. J Comp Pathol. 2010;142 Suppl 1:S22-32.

This article is based on findings from comparative epidemiology, which frequently compares human and animal populations. The author argues that meaningful comparisons between humans and animals can only be made by undertaking life span adjustment and age adjustment on animal study data, to address differences between the two populations stemming from different ‘biological ages’ and age structures, respectively. **[1a, 2a]**

102. Thyagarajan T, Totey S, Danton MJ, et al. Genetically altered mouse models: the good, the bad, and the ugly. Crit Rev Oral Biol Med. 2003;14:154-74.

This review of the current status of genetically altered mouse models highlights the challenges of understanding complex genetic and molecular mechanisms underlying craniofacial development and disease. While the human genome program has helped to generate numerous candidate genes, few genes have been characterized for their precise in vivo functions. Because some models display an unexpected or no phenotype, controversy has arisen about the value of gene-targeting strategies. The authors argue in favor of the cautious adoption of these strategies, particularly in interpreting phenotypes in craniofacial and oral biology, where many genes have pleiotropic roles. They advocate for the use of comparative genome mapping, which could provide valuable information to match mouse and human disorders accurately, and lead to testing and developing therapies for human diseases**. [1b, 2a]**

103. Tkacs NC, Thompson HJ. From bedside to bench and back again: research issues in animal models of human disease. Biol Res Nurs. 2006;8:78-88.

This article provides a basic overview of how reliability and validity of animal models can be established, focusing on models of hypoglycemia-associated autonomic failure (HAAF). The article is intended to be an introduction to translational research for nurse researchers. As such, it does not critically assess the models. **[1a]**

104. Wall RJ, Shani M. Are animal models as good as we think? Theriogenology. 2008;69:2-9.

This article suggests that animal models may have some utility in generating hypotheses (“speculation”) and elucidating basic mechanisms but are inadequate as the basis for understanding complex mechanisms or predicting human response in clinical trials ("extrapolation"). The authors express the hope that newly available data about the human genome will reveal enough about the genetic control of physiology to justify using particular animal models for very particular questions with the desired precision. **[1a, 2a]**

105. Willing AE. Experimental models: help or hindrance. Stroke. 2009;40:S152-4.

This thought-provoking editorial takes the position that the lack of translation between the preclinical animal research and clinical benefits does not lie in the animal models, but in how we use the models and how we apply this knowledge to design of clinical trials. The author argues that we need to carefully choose a preclinical stroke model, which outcome measures to use, and when to use them. She suggests that perhaps the issue is not that all the variables in preclinical studies are controlled enough, but that they are controlled too much and therefore can never truly represent the stroke patient. Moreover, if we do not want to increase variability in our animal studies, then we need to make very careful choices of clinical population to target and when to treat them when we design our clinical trials trying to mirror the animal studies precisely. If heterogeneous populations are still used in clinical trials, then the sample size must be large enough to support this and to allow for powerful post hoc analyses of subpopulations within the sample. **[1a, 2a, 3b]**

II. Articles Retrieved But Not Annotated

Category A: Added minimally to other information (i.e., limited or redundant)

1. Greek J, Shanks N. Thoughts on animal models for human disease and treatment. J Am Vet Med Assoc. 2009;235:363; author reply 364.
2. Green S. Medical progress depends on animal models—doesn't it? J R Soc Med. 2008;101:220-1.
3. Houdebine LM. Transgenic animal models in biomedical research. Methods Mol Biol. 2007;360:163-202.
4. Legg ED, Novejarque A, Rice AS. The three ages of rat: the influence of rodent age on affective and cognitive outcome measures in peripheral neuropathic pain. Pain. 2009;144:12-13.
5. Nomura T, Katsuki M, Yokoyama M, et al. Future perspectives in the development of new animal models. Prog Clin Biol Res. 1987;229:337-53.
6. Spiers AS. Studies in animals should be more like those in humans. BMJ. 2007;334:274.
7. Suckling K. Animal research: too much faith in models clouds judgement. Nature. 2008;455:460.
8. Roberts I. Animal research. Three Rs should be registration, randomisation, and reviews (systematic). BMJ. 2001;322(7302):1604.
9. Sandercock P, Roberts I. Systematic reviews of animal experiments. Lancet. 2002;360:586.
10. Shively CA, Clarkson TB. The unique value of primate models in translational research. Nonhuman primate models of women's health: introduction and overview. Am J Primatol. 2009;71:715-21.
11. Talmadge JE. Models of metastasis in drug discovery. Methods Mol Biol. 2010;602:215-33.

12. Unger EF. All is not well in the world of translational research. J Am Coll Cardiol. 2007;50:738-40.
13. van der Worp HB, Howells DW, Sena, ES, et al. Can animal models of disease reliably inform human studies? PLoS Med. 2010 Mar; 7(3): e1000245.
14. Whiteside GT, Adedoyin A, Leventhal L. Predictive validity of animal pain models? A comparison of the pharmacokinetic-pharmacodynamic relationship for pain drugs in rats and humans. Neuropharmacology. 2008;54:767-75.
15. Zhou JR, Blackburn GL. Bridging animal and human studies: what are the missing segments in dietary fat and prostate cancer? Am J Clin Nutr. 1997;66:1572S-80S.

Category B: Focused entirely on mechanisms of disease, with no attention to mechanisms of therapeutic interventions

1. Hinton DE, Hardman RC, Kullman SW, et al. Aquatic animal models of human disease: selected papers and recommendations from the 4th Conference. Comp Biochem Physiol C Toxicol Pharmacol. 2009;149:121-8.
2. Ingham PW. The power of the zebrafish for disease analysis. Hum Mol Genet. 2009;18:R107-12.
3. McMullen S, Mostyn A. Animal models for the study of the developmental origins of health and disease. Proc Nutr Soc. 2009;68:306-20.
4. Semsarian C. Use of mouse models for the analysis of human disease. Curr Protoc Hum Genet. 2002;Chapter 15:Unit 15.2.

Category C: Article could not be located

1. Rivas MA, Vecino E. Animal models and different therapies for treatment of retinitis pigmentosa. Histol Histopathol. 2009;24:1295-1322. **[C]**

III. Mapping articles onto conceptual framework

1. Strength of evidence for existence of intervention’s pathway
    a. Quality (design and execution) and strength (quantitative effect) of experimental evidence in preclinical models.
        - Bath (2009)
        - Bolton (2007)
        - Bonjour (1999)
        - Bracken (2009a, 2009b)
        - Crossley (2008)
        - Dirnagl (2009, 2006)
        - Ferrante (2009)
        - Fisher (2007)
        - Friese (2006)
        - Geerts (2009)
        - Hackam (2007, 2006)
        - Hersch (2004)
        - Horrobin (2003)

- Hsu (1993)
- Jeffery (2008)
- Kamat (2008)
- Knight (2008, 2007)
- Ledford (2008)
- Lemon (2005)
- Linder (2006)
- Lindner (2007)
- Lowenstein (2009)
- Macleod (2009, 2008, 2005, 2004)
- Malkesman (2009)
- Mao (2009)
- Markou (2009)
- Marshall (2005)
- Matthews (2008)
- Mignini (2006)
- Mogil (2009)
- O'Collins (2006)
- Pegram (2006)
- Perel (2007)
- Peters (2006)
- Philip (2009)
- Pienta (2008)
- Piper (1996)
- Pound (2004)
- Rice (2008)
- Roberts (2002)
- Roep (2007, 2004)
- Schnabel (2008)
- Sena (2007)
- Shanks (2009)
- Soubret (2009)
- Thannickal (2010)
- Thrusfield (2010)
- Tkacs (2006)
- Wall (2008)
- Willing (2009)

Number of experimental models

- Alonso de Lecinana (2001)
- Anderson (2006)
- Dyson (2009)
- Ganter (2008)
- Gold (2006)

- Gurwitz (2001)
- Hausheer (2003)
- Insel (2007)
- Loscher (2009)
- Lowenstein (2009)
- Malkesman (2009)
- Manger (2008)
- Miczek (2008)
- Opal (2009)
- Segalat (2007)
- Swanson (2004)
- Thyagarajan (2003)

b. Variety of experimental models (e.g., animal species)

- ’t Hart (2004)
- Bailey (2009)
- Baker (2008)
- Belser (2009)
- Bergman (2009)
- Bodewes (2010)
- Chatzigeorgiou (2009)
- Corry (2006)
- Dehoux (2007)
- Dixon (2009)
- Gallegos (2005)
- Hein (2003)
- Herodin (2005)
- Hersch (2004)
- Joers (2009)
- Kirschvink (2008)
- Manto (2009a, 2009b)
- Mitchell (2009)
- Muschler (2010)
- Pacharinsak (2008)
- Palena (2006)
- Pienta (2008)
- Pound (2004)
- Roep (2007, 2004)
- Schook (2005)
- Suzuki (2009)
- Swanson (2004)
- Thannickal (2010)

2. Strength of evidence that the pathway exists in human disease states.

a. Strength of evidence for animal/in vitro model’s relevance for human disease state.

- ’t Hart (2004)
- Alonso de Lecinana (2001)

- Anderson (2006)
- Bath (2009)
- Belser (2009)
- Bergman (2009)
- Bodewes (2010)
- Bonjour (1999)
- Corry (2006)
- Dehoux (2007)
- Dyson (2009)
- Ganter (2008)
- Gurwitz (2001)
- Hersch (2004)
- Horrobin (2003)
- Insel (2007)
- Jeffery (2008)
- Kamat (2008)
- Knight (2008, 2007)
- Lynch (2009)
- Manger (2008)
- Manto (2009a, 2009b)
- Markou (2009)
- Miczek (2008)
- Mitchell (2009)
- Mogil (2009)
- O'Collins (2006)
- Opal (2009)
- Pacharinsak (2008)
- Palena (2006)
- Pienta (2008)
- Pound (2004)
- Rice (2008)
- Ritter (2009)
- Roep (2007, 2004)
- Schnabel (2008)
- Schook (2005)
- Segalat (2007)
- Suzuki (2009)
- Swanson (2004)
- Thannickal (2010)
- Thrusfield (2010)
- Thyagarajan (2003)
- Wall (2008)
- Willing (2009)

b. Ex vivo evidence
   - Dragunow (2008)

- Ganter (2008)
- Hausheer (2003)
- Thannickal (2010)

c. Evidence that pathway occurs in complete physiologic system (e.g., functioning hearts vs. heart tissue.)
   - Dehoux (2007)

d. Evidence from human physiologic experiments.
   - Pacharinsak (2008)

3. Completeness of proposed mechanistic pathway. (From intervention to clinical endpoint)

   a. Gaps in pathway (including whether intervention/exposure can exert effect on target due to issues of bioavailability, metabolism, delivery, etc.)
   - Alonso de Lecinana (2001)
   - Anderson (2006)
   - Baker (2008)
   - Bergman (2009)
   - Bodewes (2010)
   - Bracken (2009a, 2009b)
   - Corry (2006)
   - Dehoux (2007)
   - DiBernardo (2006)
   - Dirnagl (2009, 2006)
   - Dixon (2009)
   - Friese (2006)
   - Geerts (2009)
   - Gold (2006)
   - Jeffery (2008)
   - Joers (2009)
   - Linder (2006)
   - Lowenstein (2009)
   - Malkesman (2009)
   - Mao (2009)
   - Mogil (2009)
   - Muschler (2010)
   - Opal (2009)
   - Pacharinsak (2008)
   - Palena (2006)
   - Pegram (2006)
   - Pienta (2008)
   - Ritter (2009)
   - Swanson (2004)
   - Thannickal (2010)

   b. Remoteness of the mechanistic outcomes from clinical outcomes.
   - ’t Hart (2004)
   - Anderson (2006)

- Corry (2006)
- DiBernardo (2006)
- Dirnagl (2009, 2006)
- Gallegos (2005)
- Gold (2006)
- Hersch (2004)
- Kamat (2008)
- Marshall (2005)
- Miczek (2008)
- Muschler (2010)
- Rice (2008)
- Soubret (2009)
- Willing (2009)

c. Strength of evidence linking proximal (i.e., surrogate) to distal (i.e., definitive) clinical endpoints
   - Bonjour (1999)
   - Dyson (2009)
   - Fisher (2007)
   - Pound (2004)

4. Evidence for alternate, competing or compensatory pathways that can:
   a. Produce outcome through pathways independent of intervention's effect
      - Dragunow (2008)
      - Ganter (2008)
      - Hausheer (2003)

   b. Produce nontherapeutic outcomes through pathways dependent on intervention
   c. Interfere with intervention's pathways

5. Strength of evidence that mechanism is similar to other interventions with known clinical effects

6. Adverse effect mechanisms
   - Bailey (2009)
   - Bergman (2009)
   - Fielden (2008)
   - Gallegos (2005)
   - Joers (2009)
   - Loscher (2009)
   - Marshall (2005)
   - Pegram (2006)
   - Soubret (2009)

Appendix A References

’t Hart BA, Amor S, Jonker M. Evaluating the validity of animal models for research into therapies for immune-based disorders. Drug Discov Today. 2004;9:517-24.

Alonso de Lecinana M, Diez-Tejedor E, Carceller F, Roda JM. Cerebral ischemia: from animal studies to clinical practice. Should the methods be reviewed? Cerebrovasc Dis. 2001;11 Suppl 1:20-30.

Anderson LM. Environmental genotoxicants/carcinogens and childhood cancer: bridgeable gaps in scientific knowledge. Mutat Res. 2006;608:136-56.

Ayhan Y, Sawa A, Ross CA, et al. Animal models of gene-environment interactions in schizophrenia. Behav Brain Res. 2009;204:274-81.

Bailey GP, Marien D. What have we learned from pre-clinical juvenile toxicity studies? Reprod Toxicol. 2009;28:226-9.

Baker DH. Animal models in nutrition research. J Nutr. 2008;138:391-6.

Bath PM, Macleod MR, Green AR. Emulating multicentre clinical stroke trials: a new paradigm for studying novel interventions in experimental models of stroke. Int J Stroke. 2009;4:471-9.

Belser JA, Szretter KJ, Katz JM, et al. Use of animal models to understand the pandemic potential of highly pathogenic avian influenza viruses. Adv Virus Res. 2009;73:55-97.

Benatar M. Lost in translation: treatment trials in the SOD1 mouse and in human ALS. Neurobiol Dis. 2007;26:1-13.

Bergman KL. The animal rule and emerging infections: the role of clinical pharmacology in determining an effective dose. Clin Pharmacol Ther. 2009;86:328-31.

Bodewes R, Rimmelzwaan GF, Osterhaus AD. Animal models for the preclinical evaluation of candidate influenza vaccines. Expert Rev Vaccines. 2010;9:59-72.

Bolton C. The translation of drug efficacy from in vivo models to human disease with special reference to experimental autoimmune encephalomyelitis and multiple sclerosis. Inflammopharmacology. 2007;15:183-7.

Bonjour JP, Ammann P, Rizzoli R. Importance of preclinical studies in the development of drugs for treatment of osteoporosis: a review related to the 1998 WHO guidelines. Osteoporos Int. 1999;9:379-93.

Bracken MB. Why animal studies are often poor predictors of human reactions to exposure. J R Soc Med. 2009;102:120-2.

Bracken MB. Why are so many epidemiology associations inflated or wrong? Does poorly conducted animal research suggest implausible hypotheses? Ann Epidemiol. 2009;19:220-4.

Chatzigeorgiou A, Halapas A, Kalafatakis K. The use of animal models in the study of diabetes mellitus. In Vivo. 2009;23:245-58.

Corry DB, Irvin CG. Promise and pitfalls in animal-based asthma research: building a better mousetrap. Immunol Res. 2006;35:279-94.

Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. Stroke. 2008;39:929-34.

Dehoux JP, Gianello P. The importance of large animal models in transplantation. Front Biosci. 2007;12:4864-80.

DiBernardo AB, Cudkowicz ME. Translating preclinical insights into effective human trials in ALS. Biochim Biophys Acta. 2006;1762:1139-49.

Dirnagl U. Bench to bedside: the quest for quality in experimental stroke research. J Cereb Blood Flow Metab. 2006;26:1465-78.

Dirnagl U, Macleod MR. Stroke research at a road block: the streets from adversity should be paved with meta-analysis and good laboratory practice. Br J Pharmacol. 2009;157:1154-6.

Dixon JA, Spinale FG. Large animal models of heart failure: a critical link in the translation of basic science to clinical practice. Circ Heart Fail. 2009;2:262-71.

Dragunow M. The adult human brain in preclinical drug development. Nat Rev Drug Discov. 2008;7:659-66.

Dyson A, Singer M. Animal models of sepsis: why does preclinical efficacy fail to translate to the clinical setting? Crit Care Med. 2009;37:S30-7.

Ferrante RJ. Mouse models of Huntington's disease and methodological considerations for therapeutic trials. Biochim Biophys Acta. 2009;1792:506-20.

Fielden MR, Kolaja KL. The role of early in vivo toxicity testing in drug discovery toxicology. Expert Opin Drug Saf. 2008;7:107-10.

Fisher M, Henninger N. Translational research in stroke: taking advances in the pathophysiology and treatment of stroke from the experimental setting to clinical trials. Curr Neurol Neurosci Rep. 2007;7:35-41.

Friese MA, Montalban X, Willcox N, et al. The value of animal models for drug development in multiple sclerosis. Brain. 2006;129:1940-52.

Gallegos RP, Nockel PJ, Rivard AL, et al. The current state of in-vivo pre-clinical animal models for heart valve evaluation. J Heart Valve Dis. 2005;14:423-32.

Ganter B, Giroux CN. Emerging applications of network and pathway analysis in drug discovery and development. Curr Opin Drug Discov Devel. 2008;11:86-94.

Geerts H. Of mice and men: bridging the translational disconnect in CNS drug discovery. CNS Drugs. 2009;23:915-26.

Gold R, Linington C, Lassmann H. Understanding pathogenesis and therapy of multiple sclerosis via animal models: 70 years of merits and culprits in experimental autoimmune encephalomyelitis research. Brain. 2006;129:1953-71.

Greek J, Shanks N. Thoughts on animal models for human disease and treatment. J Am Vet Med Assoc. 2009;235:363; author reply 364.

Green S. Medical progress depends on animal models—doesn't it? J R Soc Med. 2008;101:220-1.

Gurwitz D, Weizman A. Animal models and human genome diversity: the pitfalls of inbred mice. Drug Discov Today. 2001;6:766-8.

Hackam DG. Translating animal research into clinical benefit. BMJ. 2007;334:163-4.

Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. JAMA. 2006;296:1731-2.

Hausheer FH, Kochat H, Parker AR, et al. New approaches to drug discovery and development: a mechanism-based approach to pharmaceutical research and its application to BNP7787, a novel chemoprotective agent. Cancer Chemother Pharmacol. 2003;52 Suppl 1:S3-15.

Hein WR, Griebel PJ. A road less traveled: large animal models in immunological research. Nat Rev Immunol. 2003;3:79-84.

Herodin F, Thullier P, Garin D, et al. Nonhuman primates are relevant models for research in hematology, immunology and virology. Eur Cytokine Netw. 2005;16:104-16.

Hersch SM, Ferrante RJ. Translating therapies for Huntington's disease from genetic animal models to clinical trials. NeuroRx. 2004;1:298-306.

Hinton DE, Hardman RC, Kullman SW, et al. Aquatic animal models of human disease: selected papers and recommendations from the 4th Conference. Comp Biochem Physiol C Toxicol Pharmacol. 2009;149:121-8.

Horrobin DF. Modern biomedical research: an internally self-consistent universe with little contact with medical reality? Nat Rev Drug Discov. 2003; 2(2):151-4.

Houdebine LM. Transgenic animal models in biomedical research. Methods Mol Biol. 2007;360:163-202.

Hsu CY. Criteria for valid preclinical trials using animal stroke models. Stroke. 1993 May; 24(5): 633-6.

Ingham PW. The power of the zebrafish for disease analysis. Hum Mol Genet. 2009;18:R107-12.

Insel TR. From animal models to model animals. Biol Psychiatry. 2007;62:1337-9.

Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005; 294:218-228.

Jeffery EH, Keck AS. Translating knowledge generated by epidemiological and in vitro studies into dietary cancer prevention. Mol Nutr Food Res. 2008;52 Suppl 1:S7-17.

Joers VL, Emborg ME. Preclinical assessment of stem cell therapies for neurological diseases. ILAR J. 2009;51:24-41.

Kamat CD, Gadal S, Mhatre M, et al. Antioxidants in central nervous system diseases: preclinical promise and translational challenges. J Alzheimers Dis. 2008;15:473-93.

Kirschvink N, Reinhold P. Use of alternative animals as asthma models. Curr Drug Targets. 2008;9:470-84.

Knight A. Animal experiments scrutinised: systematic reviews demonstrate poor human clinical and toxicological utility. ALTEX. 2007;24:320-5.

Knight A. Reviewing existing knowledge prior to conducting animal studies. Altern Lab Anim. 2008 Dec; 36(6): 709-12.

Knight A. Systematic reviews of animal experiments demonstrate poor contributions toward human health care. Rev Recent Clin Trials. 2008;3:89-96.

le Coutre P, Mologni L, Cleris L, Marchesi E, Buchdunger E, Giardini R, Formelli F, Gambacorti-Passerini C. In vivo eradication of human BCR/ABL-positive leukemia cells with an ABL kinase inhibitor. J Natl Cancer Inst. 1999 20;91:163-168.

Ledford H. Translational research: the full cycle. Nature. 2008;453:843-5.

Legg ED, Novejarque A, Rice AS. The three ages of rat: the influence of rodent age on affective and cognitive outcome measures in peripheral neuropathic pain. Pain. 2009;144:12-13.

Lemon R, Dunnett SB. Surveying the literature from animal experiments. BMJ. 2005;330:977-8.

Linder S, Shoshan MC. Is translational research compatible with preclinical publication strategies? Radiat Oncol. 2006;1:4.

Lindner MD. Clinical attrition due to biased preclinical assessments of potential efficacy. Pharmacol Ther. 2007;115:148-75.

Loscher W. Preclinical assessment of proconvulsant drug activity and its relevance for predicting adverse events in humans. Eur J Pharmacol. 2009;610:1-11.

Lowenstein PR, Castro MG. Uncertainty in the translation of preclinical experiments to clinical trials: Why do most Phase III clinical trials fail? Current Gene Therapy. 2009; 9: 368-74.

Lynch VJ. Use with caution: developmental systems divergence and potential pitfalls of animal models. Yale J Biol Med. 2009;82: 53-66.

Macleod MR, Ebrahim S, Roberts I. Surveying the literature from animal experiments: systematic review and meta-analysis are important contributions. BMJ. 2005;331:110.

Macleod MR, Fisher M, O’Collins V, et al. Good laboratory practice: preventing introduction of bias at the bench. Stroke. 2009;40:e50-2.

Macleod MR, O’Collins T, Howells DW, et al. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke. 2004;35:1203-8.

Macleod MR, van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke. 2008;39:2824-9.

Malkesman O, Austin DR, Chen G, et al. Reverse translational strategies for developing animal models of bipolar disorder. Dis Model Mech. 2009;2:238-45.

Manger PR, Cort J, Ebrahim N, et al. Is 21st century neuroscience too focused on the rat/mouse model of brain function and dysfunction? Front Neuroanat. 2008;2:5.

Manto M, Marmolino D. Animal models of human cerebellar ataxias: a cornerstone for the therapies of the twenty-first century. Cerebellum. 2009;8:137-54.

Manto M, Marmolino D. Cerebellar disorders—at the crossroad of molecular pathways and diagnosis. Cerebellum. 2009;8:417-22.

Mao J. Translational pain research: achievements and challenges. J Pain. 2009;10:1001-11.

Markou A, Chiamulera C, Geyer MA, et al. Removing obstacles in neuroscience drug discovery: the future path for animal models. Neuropsychopharmacology. 2009; 34:74-89.

Marshall JC, Deitch E, Moldawer LL, et al. Preclinical models of shock and sepsis: what can they tell us? Shock. 2005;24 Suppl 1:1-6.

Matthews RA. Medical progress depends on animal models—doesn’t it? J R Soc Med. 2008;101:95-8.

McMullen S, Mostyn A. Animal models for the study of the developmental origins of health and disease. Proc Nutr Soc. 2009;68:306-20.

Miczek KA, de Wit H. Challenges for translational psychopharmacology research—some basic principles. Psychopharmacology (Berl). 2008;199:291-301.

Mignini LE, Khan KS. Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. BMC Med Res Methodol. 2006;6:10.

Mitchell BF, Taggart MJ. Are animal models relevant to key aspects of human parturition? Am J Physiol Regul Integr Comp Physiol. 2009;297:R525-45.

Mogil JS. Animal models of pain: progress and challenges. Nat Rev Neurosci. 2009;10:283-94.

Muschler GF, Raut VP, Patterson TE, et al. The design and use of animal models for translational research in bone tissue engineering and regenerative medicine. Tissue Eng Part B Rev. 2010 Feb;16(1):123-45.

Nomura T, Katsuki M, Yokoyama M, et al. Future perspectives in the development of new animal models. Prog Clin Biol Res. 1987;229:337-53.

O'Collins VE, Macleod MR, Donnan GA, et al. 1,026 experimental treatments in acute stroke. Ann Neurol. 2006;59:467-77.

Opal SM, Patrozou E. Translational research in the development of novel sepsis therapeutics: logical deductive reasoning or mission impossible? Crit Care Med. 2009;37:S10-5.

Pacharinsak C, Beitz A. Animal models of cancer pain. Comp Med. 2008;58:220-33.

Palena C, Abrams SI, Schlom J, et al. Cancer vaccines: preclinical studies and novel strategies. Adv Cancer Res. 2006;95:115-45.

Pegram M, Ngo D. Application and potential limitations of animal models utilized in the development of trastuzumab (Herceptin): a case study. Adv Drug Deliv Rev. 2006;58:723-34.

Perel P, Roberts I, Sena E, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ. 2007;334:197.

Peters JL, Sutton AJ, Jones DR, et al. A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. J Environ Sci Health B. 2006;41:1245-58.

Philip M, Benatar M, Fisher M, et al. Methodological quality of animal studies of neuroprotective agents currently in phase II/III acute ischemic stroke trials. Stroke. 2009;40:577-81.

Pienta KJ, Abate-Shen C, Agus DB, et al. The current state of preclinical prostate cancer animal models. Prostate. 2008;68:629-39.

Piper RD, Cook DJ, Bone RC, et al. Introducing Critical Appraisal to studies of animal models investigating novel therapies in sepsis. Crit Care Med. 1996;24:2059-70.

Pound P, Ebrahim S, Sandercock P, et al. Where is the evidence that animal research benefits humans? BMJ. 2004;328:514-7.

Rice AS, Cimino-Brown D, Eisenach JC, et al. Animal models and the prediction of efficacy in clinical trials of analgesic drugs: a critical appraisal and call for uniform reporting standards. Pain. 2008;139:243-7.

Ritter T, Nosov M, Griffin MD. Gene therapy in transplantation: Toward clinical trials. Curr Opin Mol Ther. 2009;11:504-12.

Rivas MA, Vecino E. Animal models and different therapies for treatment of retinitis pigmentosa. Histol Histopathol. 2009;24:1295-1322.

Roberts I. Animal research. Three Rs should be registration randomisation, and reviews (systematic). BMJ. 2001;322(7302):1604.

Roberts I, Kwan I, Evans P, et al. Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. BMJ. 2002;324:474-6.

Roep BO. Are insights gained from NOD mice sufficient to guide clinical translation? Another inconvenient truth. Ann N Y Acad Sci 2007 Apr 1103: 1-10.

Roep BO, Atkinson M, von Herrath M. Satisfaction (not) guaranteed: re-evaluating the use of animal models of type 1 diabetes. *Nat Rev Immunol*. 2004 Dec;4(12):989-97.

Rosenblum WI. Criteria for valid preclinical trials using animal stroke models. Stroke. 1993;24:1601-2.

Sandercock P, Roberts I. Systematic reviews of animal experiments. Lancet. 2002;360:586.

Schnabel J. Neuroscience: Standard model. Nature. 2008;454:682-5.

Schook L, Beattie C, Beever J, et al. Swine in biomedical research: creating the building blocks of animal models. Anim Biotechnol. 2005;16:183-90.

Segalat L. Invertebrate animal models of diseases as screening tools in drug discovery. ACS Chem Biol. 2007;2:231-6.

Semsarian C. Use of mouse models for the analysis of human disease. Curr Protoc Hum Genet. 2002; Chapter 15:Unit 15.2.

Sena E, van der Worp HB, Howells D, et al. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci. 2007;30:433-9.

Shanks N, Greek R, Greek J. Are animal models predictive for humans? Philos Ethics Humanit Med. 2009;4:2.

Shively CA, Clarkson TB. The unique value of primate models in translational research. Nonhuman primate models of women's health: introduction and overview. Am J Primatol. 2009;71:715-21.

Soubret A, Helmlinger G, Dumotier B, et al. Modeling and simulation of preclinical cardiac safety: towards an integrative framework. Drug Metab Pharmacokinet. 2009;24:76-90.

Spiers AS. Studies in animals should be more like those in humans. BMJ. 2007;334:274.

Suckling K. Animal research: too much faith in models clouds judgement. Nature. 2008;455:460.

Suzuki Y, Yeung AC, Ikeno F. The pre-clinical animal model in the translational research of interventional cardiology. JACC Cardiovasc Interv. 2009;2:373-83.

Swanson KS, Mazur MJ, Vashisht K, et al. Genomics and clinical medicine: rationale for creating and effectively evaluating animal models. Exp Biol Med (Maywood). 2004;229:866-75.

Talmadge JE. Models of metastasis in drug discovery. Methods Mol Biol. 2010;602:215-33.

Thannickal VJ, Roman J. Challenges in translating preclinical studies to effective drug therapies in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2010 Mar:181(6):532-3.

Thrusfield MV. Ageing in animal populations—an epidemiological perspective. J Comp Pathol. 2010;142 Suppl 1:S22-32.

Thyagarajan T, Totey S, Danton MJ, et al. Genetically altered mouse models: the good, the bad, and the ugly. Crit Rev Oral Biol Med. 2003;14:154-74.

Tkacs NC, Thompson HJ. From bedside to bench and back again: research issues in animal models of human disease. Biol Res Nurs. 2006;8:78-88.

Unger EF. All is not well in the world of translational research. J Am Coll Cardiol. 2007;50:738-40.

van der Worp HB, Howells DW, Sena ES, et al. Can animal models of disease reliably inform human studies? PLoS Med. 2010 Mar; 7(3): e1000245.

Wall RJ, Shani M. Are animal models as good as we think? Theriogenology. 2008;69:2-9.

Whiteside GT, Adedoyin A, Leventhal L. Predictive validity of animal pain models? A comparison of the pharmacokinetic-pharmacodynamic relationship for pain drugs in rats and humans. Neuropharmacology. 2008;54:767-75.

Willing AE. Experimental models: help or hindrance. Stroke. 2009;40:S152-4.

Zhou JR, Blackburn GL. Bridging animal and human studies: what are the missing segments in dietary fat and prostate cancer? Am J Clin Nutr. 1997;66:1572S-80S.

Appendix B. Annotated Bibliography of Surrogate Endpoints Literature, With Framework Mapping

This component of the project was conducted before the framework described in this document was fully developed and finalized. A preliminary framework was used into which the various articles were mapped. This mapping was done in the form of bolded codes that appear at the end of each article description, and correspond to the following dimensions.

1) **Strength of evidence for existence of intervention's pathway**
   a) Quality (design and execution) and strength (quantitative effect) of experimental evidence in preclinical models.
   b) Number of experimental models
   c) Variety of experimental models (e.g. animal species)
2) **Strength of evidence that the pathway exists in human disease states.**
   a) Strength of evidence for animal/in vitro model's relevance for human disease state.
   b) Ex vivo evidence
   c) Evidence that pathway occurs in complete physiologic system (e.g. functioning hearts vs. heart tissue.)
   d) Evidence from human physiologic experiments.
3) **Completeness of proposed mechanistic pathway.** (From intervention to clinical endpoint)
   a) Gaps in pathway (including whether intervention/exposure can exert effect on target due to issues of bioavailability, metabolism, delivery, etc.)
   b) Remoteness of the mechanistic outcomes from clinical outcomes.
   c) Strength of evidence linking proximal (i.e. surrogate) to distal (i.e. definitive) clinical endpoints
4) **Evidence for alternate, competing or compensatory pathways that can:**
   a) Produce outcome through pathways independent of intervention's effect
   b) Produce non-therapeutic outcomes through pathways dependent on intervention
   c) Interfere with intervention's pathways
5) **Strength of evidence that mechanism is similar to other interventions with known clinical effects**
6) **Adverse event mechanisms**

I. Annotated articles

1. Altar CA, Bounos D Amakye D, et al. A prototypical process for creating evidentiary standards for biomarkers and diagnostics. Clin Pharmacol Ther. 2008 Feb; 83(2): 368-71.
2. Altar CA. The Biomarkers Consortium: on the critical path of drug discovery. Clin Pharmacol Ther. 2008 Feb; 83(2): 361-4.

The above two articles present a framework for assessing evidence for biomarker qualification, a task viewed as important to the FDA's Critical Path initiative. Table 1 of the Altar et al. paper includes a "Prototype 'evidence map,'" which contains categorical description of different types of scientific evidence potentially relevant to biomarker qualification. It outlines a system of assigned letter grades to subcategories of evidence. "Theory on biologic plausibility" is weighted the least important of the seven subcategories, followed by "Interaction with pharmacologic target," and "Pharmacologic mechanistic response." This might serve as a useful model for BMEBM instrument development. **[1a, 3c]**

3. Alymani NA, Smith MD, Williams DJ, et al. Predictive biomarkers for personalised anti-cancer drug use: Discovery to clinical implementation. Eur J Cancer. 2010 Mar;46(5):869-79.

This article examines failure to translate initially promising cancer (especially solid tumor) predictive biomarkers into clinically useful applications and highlights the need to develop a robust clinical biomarker development methodology: discovery, validation, qualification and implementation; see Figure 1. Most useful to BMEBM is discussion of discovery and qualification. Due to molecular complexity and heterogeneity, there is a dearth of evidence at discovery phase (re: causal mechanistic relationship between a particular molecular pathway and the clinical outcome in individual patients). Frequently, correlative relationships between biomarkers and clinical endpoints are relied upon to identify leads, especially using ‘omic’ platforms. Qualification (defined as “the evidentiary process of establishing a causal or correlative relationship between the biomarker and the clinical end-point or other biological or pathological end-point”) relies in part on appeals to expert opinion (panels convened by the FDA). **[1a, 3a, 3c]**

4. Antoine DJ, Mercer AE, Williams DP, et al. Mechanism-based bioanalysis and biomarkers for hepatic chemical stress. Xenobiotica. 2009;39:565-77.

This review article summarizes the potential of novel mechanism-based biomarkers of hepatic stress, which provide information concerning the molecular basis of drug-induced liver injury (DILI). The authors emphasize the importance of our ability to link the chemistry of the drug to a clinically observed adverse drug reaction by an understanding of the intracellular and extracellular signaling pathways involved (that is, the mechanisms of action). They discuss a number of examples/models of DILI biomarkers, endorsing the view that an integrated analysis of the biochemical, molecular, and cellular events provides an understanding of biological factors which ultimately determine the balance between xenobiotic detoxification and liver injury. **[1b, 3a, 6]**

5. Bhattacharya S, Mariani TJ. Array of hope: expression profiling identifies disease biomarkers and mechanism. Biochem Soc Trans. 2009;37:855-62.

This article examines the utility of genome-wide microarray technologies in the identification of biomarkers and disease mechanisms. The authors—reviewing recent advances in the area of respiratory diseases — classify microarray-based discovery into three components: biomarker detection, disease (sub)classification and identification of causal mechanism. They describe initial limitations/deficiencies in using microarray, including: experimental design/study size, analytical methods and probe sequences. As many of these limitations are better understood or have been overcome, they strike a hopeful tone that microarray, when used in combination with animal models and genetic studies, particularly focusing on quantitative variable analysis, can provide unexpected power to identify disease mechanisms. **[2b, 3c]**

6. Bhogal N, Balls M. Translation of new technologies: from basic research to drug discovery and development. Curr Drug Discov Technol. 2008;5:250-62.

This article describes new omics’ technologies, high-throughput analyses and imaging techniques that have potentially enormous value to the drug development process. These are leading to a gradual shift away from the traditional animal testing-dominated paradigm, to one based on using human cells and tissues alongside human microdosing to develop new therapeutic agents. The authors argue that a human-focused approach holds the key to reducing the attrition rates in drug development, particularly with regard to the development of treatments for human diseases with complex and varied etiologies that cannot readily be simulated using animal models or which are to be targeted by highly human-specific agents. Tables 1 and 2, and Figures 2 and 3 are particularly useful for BMEBM purposes. **[2b, 3c]**

7. Biomarker Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar; 69(3): 89-95.

This commentary by an expert group convened by the National Institutes of Health provides definitions of the terms biomarker, clinical endpoint and surrogate endpoint in order to help build consensus in describing biological measurements used in therapeutic development and assessment. Most useful for BMEBM is the conceptual model of the relationship of biomarkers, surrogate endpoints, and the process of evaluating therapeutic interventions, which includes a box labeled “Evidence that a biomarker is reasonably likely to predict clinical benefit or risk (Figure 1). Also includes very brief discussion of biomarker validation as reflecting causal or mechanistic associations of the intervention with the disease process (as opposed to a statistical approach). **[3b, 3c, 5]**

8. Boffetta P. Biomarkers in cancer epidemiology: an integrative approach. Carcinogenesis. 2010;31:121-6.

This article identifies various reasons for the increased use of biomarkers in cancer epidemiology, including: (1) the fact that the identification of new carcinogens, characterized by complex exposure circumstances and weak effects, has become increasingly difficult with traditional epidemiological approaches; (2) the increasing understanding of mechanisms of carcinogenesis and (3) technical developments in molecular biology and genetics. The author distinguishes between biomarkers of exposure, effect (biological events that take place in the continuum between exposure and cancer development), and susceptibility (identification of high-risk subgroups of the population), arguing that molecular epidemiology should strive to address several components of the carcinogenic process in a single conceptual model. While this article does not focus on biomarkers in the context of therapeutic interventions, the discussion of effect biomarkers—particularly how they may be used to increase the specificity and the sensitivity in the definition of the outcome—is relevant to BMEBM. **[2a, 3a, 3c]**

9. Carden CP, Banerji U, Kaye SB, et al. From darkness to light with biomarkers in early clinical trials of cancer drugs. Clin Pharmacol Ther. 2009;85:131-3.

This opinion piece suggests that the traditional clinical trial infrastructure is inefficient, slow and costly, and serves the cancer research community particularly poorly. The authors recommend more widespread use of adaptive early biomarker-driven clinical trials testing molecularly targeted agents that are able to not only question but also answer key scientific and clinical hypotheses. Such trials can arguably have a major impact on understanding of disease biology, decreasing the risk of late and costly drug attrition. These trials of targeted agents require definitive evidence of target modulation in tumor cells by the agent under evaluation, in carefully selected patients. **[3a]**

10. Carroll KJ. Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology. Pharm Stat. 2007;6:253-60.

In this personal reflection piece, the author sounds a cautionary note concerning the use of biomarkers in drug development. Where biomarkers are used for candidate drug screening for intrinsic activity or proof of mechanism, there seems little impediment to their use and the burden falls squarely on the sponsor to be sure the biomarker endpoint helps to make the right decisions. However, using biomarkers as surrogate endpoints for clinical outcome to support drug approval is more troublesome. Establishing a new biomarker as a true surrogate endpoint using published statistical criteria is extremely demanding, if not impossible. The author argues that a lower burden of evidence is required and, consequently, that greater risks be taken, in order to use new biomarkers as substitutes for clinical outcome. **[3b, 3c]**

11. Chau CH, Rixe O, McLeod H, et al. Validation of analytic methods for biomarkers used in drug development. Clin Cancer Res. 2008;14:5967-76.

This paper focuses on the general principles of biomarker validation in the drug development process, with an emphasis on assay validation. Table 1 describes the potential uses of biomarkers in each phase of drug development, and includes a “preclinical studies” phase (development of appropriate animal models that feature biomarker properties comparable with those seen in patient populations to enhance predictivity; role in validation of new disease models; assess toxicity and safety of drug). **[2b, 3a]**

12. Chetty RK, Ozer JS, Lanevschi A, et al. A systematic approach to preclinical and clinical safety biomarker qualification incorporating Bradford Hill’s principles of causality association. Clin Pharmacol Ther. 2010 Aug;88(2):260-2.

This article addresses questions regarding the optimal methods of collecting and evaluating scientific evidence for the clinical qualification of a biomarker, with a focus on toxicology biomarkers. The authors propose a novel application to assist and accelerate the drug development process by prioritizing biomarker candidates and evidence, an application based on Bradford Hill’s principles of causality association. The criteria related to biological plausibility and coherence are of interest for BMEBM purposes. **[3a, 6]**

13. Clark DP. Ex vivo biomarkers: functional tools to guide targeted drug development and therapy. Expert Rev Mol Diagn. 2009;9:787-94.

In this review, the authors explore the promise of ex vivo biomarkers in the development of cancer drugs. Most of the currently utilized predictive biomarkers for therapeutic decision-making provide information regarding the presence or absence of the drug target but reveal little about the functional circuitry of the signaling network that the drug must also impact. Ex vivo biomarkers are dynamic molecular markers evoked from living tumor cells after removal from the patient. Live tumor cells are procured from a patient and are exposed to a stimulus, such as a growth factor, sometimes in the presence of a modulator like a drug, to evoke ex vivo biomarkers. These biomarkers are then assembled into a signaling profile that provides functional information about the tumor that is not available using traditional processing. The authors argue that ex vivo biomarkers provide valuable mechanistic information that may facilitate drug development and guide the clinical selection of targeted therapeutics or identify potential responder subpopulations (e.g., ex vivo biomarkers from model systems, such as murine xenografts). **[2b, 3b]**

14. Coate LE, John T, Tsao MS, et al. Molecular predictive and prognostic markers in non-small-cell lung cancer. Lancet Oncol. 2009;10:1001-10.

This paper reviews current predictive and prognostic biomarkers in non-small-cell lung cancer (NSCLC). The authors assess their potential clinical use and explore recent data pertaining to genome-wide approaches for treatment selection in NSCLC. The paper raises some interesting questions of relevance to BMEBM. For instance, even as molecular analysis advances, we still do not know whether the molecular profile of a tumor changes at the time of disease recurrence after surgery, or even after therapy for more advanced disease. There is little information as to whether primary and metastatic tumors always share the same molecular profile, although there is some evidence for molecular discordance between early and metastatic disease. If this finding is shown to be a frequent occurrence, repeat biopsy with molecular profiling of fresh tissue might be required when treatments change, especially if the new treatment has a specific molecular target. Table 3 summarizes the prognostic and predictive markers in NSCLC, and includes a “level of evidence” column. **[2b, 3a]**

15. Colburn WA. Biomarkers in drug discovery and development: from target identification through drug marketing. J Clin Pharmacol. 2003;43:329-41.
16. Colburn WA. Optimizing the use of biomarkers, surrogate endpoints, and clinical endpoints for more efficient drug development. J Clin Pharmacol. 2000;40:1419-27.

This is a highly useful pair of papers by an author who has thought deeply about biomarkers and biological mechanisms of disease progression and therapeutic intervention. The author argues that early in discovery and development, the biomarker should at least reflect activity that is mediated through the theoretical disease mechanism of action. Later in development, the biomarker should represent mechanism-based processes that are critical to disease progression and that are appropriately altered by effective therapeutic interventions. Some biomarkers, he points out, are simply associated with the disease, do not drive the disease process, or are not altered by therapeutic intervention that acts on the disease mechanism. False-positive results occur when there is an assumption that the biomarker is a critical part of the disease process when in fact it is only loosely associated with the disease process or a random event that has inadvertently coincided with disease diagnosis and progression. The author raises the interesting point that the lack of agreement between biomarkers and clinical endpoints—often attributed as the “fault” of the former—is often caused by the selection of an unreliable endpoint. **[3a, 3b, 3c]**

17. Collins CD, Purohit S, Podolsky RH, et al. The application of genomic and proteomic technologies in predictive, preventive and personalized medicine. Vascul Pharmacol. 2006;45:258-67.

This article reviews recent technological advances in genetics, genomics, proteomics, and bioinformatics that may help advance and accelerate biomarker discovery. The authors focus on the development of predictive biomarkers for chronic diseases such as diabetes, cancer and hypertension, suggesting that effective prevention requires sensitive and specific biomarkers that can accurately identify the at-risk population before the onset of clinical symptoms. **[2b, 3b]**

18. Danhof M, Alvan G, Dahl SG, et al. Mechanism-based pharmacokinetic-pharmacodynamic modeling-a new classification of biomarkers. Pharm Res. 2005;22:1432-7.

This article lays out the principles of mechanism-based PK/PD modeling, which the authors argue "constitutes a basis for better understanding of the biological system of interest providing a basis for extrapolation and prediction" than empirical descriptive models (presumably they mean correlative relationships). They propose a seven-level classification of biomarkers containing specific mathematical expressions for processes on the causal path between drug administration and response (described in Figure 1). Their "causal-chain" approach emphasizes construct validity (evidence that a biomarker shares a causal mechanism with an ultimate clinical endpoint), similar to Malkesman et al. (2009). **[2a, 3c]**

19. Danna EA, Nolan GP. Transcending the biomarker mindset: deciphering disease mechanisms at the single cell level. Curr Opin Chem Biol. 2006;10:20-7.

The authors of this insightful piece recognize the promise of proteomics for advancing our understanding and treatment of many diseases. They caution, however, that while 'omic approaches can inform us about molecular signatures as markers of disease state, continued research into biological mechanism is critical. Taking examples from viral infection, autoimmunity, and cancer research, they argue that detailed analyses of disease-associated signaling networks have the potential to be more mechanistically informative than large-scale proteomic profiling approaches, providing insight into the cellular processes involved in pathogenesis, disease progression and therapeutic resistance. **[2a, 2b, 3a]**

20. Day M, Balci F, Wan HI, et al. Cognitive endpoints as disease biomarkers: optimizing the congruency of preclinical models to the clinic. Curr Opin Investig Drugs. 2008;9:696-706.

This article examines the role of cognitive biomarkers in neuropsychiatric and neurodegenerative disorders. By focusing on cognition as a potential disease biomarker, the authors raise important questions of how symptoms of human disease are modeled in animals, including: (1) Will the preclinical and clinical cognitive endpoints track with the initiation, progression, remission and relapse of the disease?; (2) Can the same test parameters be employed across species?; (3) Has congruency between the effects of compounds been demonstrated in the same tests in rodents, non-human primates and humans?; and (4) Can the preclinical and clinical tests differentiate extraneous confounding variables? The paper includes what the authors refer to as a "utilitarian classification" of biomarker types and a tool for assessing the strength of biomarkers (see Table 1; explored in greater detail in Day et al. 2009). **[1c, 2a, 3a]**

21. Day M, Rutkowski JL, Feuerstein GZ. Translational medicine—a paradigm shift in modern drug discovery and development: the role of biomarkers. Adv Exp Med Biol. 2009;655:1-12.

The authors of this paper present a "utilitarian classification" of biomarkers. The typology consists of 5 types of biomarkers (target validation, target/compound interaction, PD activitiy, disease biomarker and disease modification, patient stratification and adaptive design). The utility of this system is represented in Figures 2-4 of the paper, which suggest a semi-quantitative scoring system that helps assess the strength of the program overall and identification of the areas of weaknesses in each of the biomarkers needed along the biological progression path. **[3a, 3b]**

22. De Gruttola VG, Clax P, DeMets DL, et al. Considerations in the evaluation of surrogate endpoints in clinical trials. Summary of a National Institutes of Health workshop. Control Clin Trials. 2001;22:485-502.

This article focuses primarily on the development of statistical models to evaluate biomarkers and surrogate endpoints. The authors recognize the importance of using emerging mechanistic knowledge to build appropriate statistical models, arguing that new techniques are needed for testing the validity of presumed mechanisms and for updating the evaluation of biomarkers and surrogates with new mechanistic evidence and clinical data. The authors make a number of recommendations in this area, including: developing models that can accommodate measurement error and missing data/informative censoring for investigating biomarkers in different disease areas; evaluating latent variable and other models using patient-specific data for prediction; building models to incorporate longitudinal measurement of biomarkers and sequential treatments; considering a variety of estimation procedures (e.g., classical methods, empirical Bayes, or Markov Chain Monte Carlo techniques). **[3c]**

23. Dhani N, Siu LL. Clinical trials and biomarker development with molecularly targeted agents and radiotherapy. Cancer Metastasis Rev. 2008;27:339-49.

This paper examines paradigms of drug development used for cytotoxic chemotherapeutic agents (CCAs). The authors state that high throughput techniques in genomic, proteomic and metabolomic profiling should allow for more effective preclinical investigation of CCAs with the identification of biomarkers or indicators of treatment response, ultimately resulting in increased clinical efficacy and appropriate patient selection. They caution that while assessment of CCAs in validated preclinical models may be helpful to guide clinical trial design and to identify patient sub-populations most likely to benefit, results from preclinical experiments may not necessarily correlate with the clinical experience. They strike a hopeful note, suggesting that the ongoing evolution of more reproducible ‘omic profiling assays will make the early evaluation of multiple markers in human phase 0 studies a useful addition to preclinical biomarker development. **[2b, 3b]**

24. Dudley JT, Butte AJ. Identification of discriminating biomarkers for human disease using integrative network biology. Pac Symp Biocomput. 2009;27-38.

This paper presents a framework for the identification of disease-specific protein biomarkers through the integration of biofluid proteomes and inter-disease genomic relationships using a network paradigm. The authors make the case that while a more traditional biomarker discovery process might start with the disease of interest to identify biomarker candidates in a “bottom-up” approach, a “top-down” approach may be more efficient. This approach begins with the broad space of human disease and full compliments of biofluid proteomes to quickly discern candidate protein biomarkers discriminately associated with a disease condition. **[2b]**

25. Fine BM, Amler L. Predictive biomarkers in the development of oncology drugs: a therapeutic industry perspective. Clin Pharmacol Ther. 2009;85:535-8.

This brief perspective piece examines the opportunities and challenges associated with codeveloping predictive biomarkers alongside new therapeutics, using a few specific examples targeting cancer. The authors make two main points: (1) In spite of the success in the development of trastuzumab/Herceptin, incorporating predictive biomarkers into drug development for faster, smaller, cheaper, and ultimately more successful clinical development is difficult. The instances in which researchers are working with the prior knowledge (incl. preclinical mechanistic evidence) that was available in the development of trastuzumab may be rare. (2) Our understanding of predictive biomarkers for a particular drug is likely to evolve for many years after initial regulatory approval is given for the drug; in many cases, highly predictive biomarkers may not be identified and confirmed until after completion of pivotal clinical studies. **[1a, 3a, 3c]**

26. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125:605-13.

This widely-cited paper begins with the observation that effects on surrogate end points often do not predict the true clinical effects of interventions. The authors examine many explanations for this failure (e.g., existence of causal pathways of the disease process that are not mediated through the surrogate end point and that might be influenced differently by the intervention). They contend that the most plausible explanation for failure is usually that the intervention has unintended mechanisms of action that are independent of the disease process. These unintended mechanisms can readily cause the effect on the true clinical outcome to be inconsistent with what would have been expected solely on the basis of evaluation of surrogate end points. **[3c]**

27. Fleming TR. Surrogate endpoints and FDA’s accelerated approval process. Health Aff (Millwood). 2005;24:67-78.

Building on Fleming and DeMets (1996), this paper considers issues related to validating surrogate endpoints (i.e., identifying when effects on biomarkers can accurately predict when treatment truly provides tangible benefit to patients). The author proposes an endpoint hierarchy describing the relative reliability of outcome measures when used to evaluate clinical benefit. He also considers the issues in the implementation of the FDA’s accelerated-approval process, where treatments only known to be biologically active can be marketed to the public while scientific trials are under way to determine whether these agents truly are more effective than toxic. Though focused on clinical as opposed to preclinical studies, the paper raises fundamental questions about evidence evaluation. The author argues that validation of a surrogate should be based on both in-depth clinical insights and empirical evidence in which one should (ideally) have a comprehensive understanding of the causal pathways of the disease process and of the intervention’s unintended and intended mechanisms of action. The paper includes a proposed 4-level endpoint hierarchy for outcome measures, and the author laments that most current surrogates occupy the lowest level, where evidence regarding biological mechanism is weakest. **[2a, 3c]**

28. Floyd E, McShane TM. Development and use of biomarkers in oncology drug development. Toxicol Pathol. 2004;32 Suppl 1:106-15.

This paper offers a useful description of the preclinical development of biomarkers in oncology research. Preclinically, biomarkers can facilitate selection of animal models and of lead compounds tested in those models. They can demonstrate pharmacological and PD mechanisms-of-action in in vitro and in vivo preclinical models. Some biomarkers such as those measuring apoptosis or signaling pathways can be used to mathematically model the effects of anticancer drug combinations to predict optimum clinical treatment regimens. Biomarker assays already established for diagnostic use and treatment monitoring in humans can be modified and evaluated preclinically to assess their validity for use in clinical trials with a particular drug candidate. The authors present a useful conceptual framework for assessing evidence of biomarker safety and efficacy (Figure 1), as well as an evidence-based 3-level hierarchy (Figure 2). Many Level-1 and Level-2 (the lower two levels) biomarkers destined for translation to the clinic are evaluated preclinically to establish that a marker is robust in a relevant model using the particular drug under development. **[1a, 2a, 3a]**

29. Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nat Rev Drug Discov. 2003;2:566-80.

This paper provides a good overview of biomarker terminology and validation, and provides useful examples from cardiology, neurology, oncology, psychiatry, as well as emerging imaging technologies. The authors provide a basic typology of biomarkers to guide drug development, consisting of target, mechanism and clinical biomarkers (Figure 2). They argue that the objective is to deploy these biomarkers as early as possible, first to confirm hitting the target and then to test whether hitting a target alters the pathophysiological mechanism and altering this mechanism affects clinical status. The authors also make use of the DeMets/Fleming diagrams (Fleming and Demets, 1996) re: how biomarkers can lead to erroneous conclusions (Figure 3). In addition to the four categories of errors discussed by DeMets/Fleming, they offer a fifth, more subtle, reason that a biomarker might ‘fail’, and this is by providing potentially misleading information. The authors contend that erroneous conclusions can be drawn regarding the usefulness of a biomarker when the biomarker might not correlate well with the gold-standard clinical assessor because the biomarker is more sensitive or the gold-standard assessor is irrelevant to a subset of the patient population, a novel mechanism, or a new indication. Finally, the authors suggest that the development of biomarkers for diagnostic and prognostic use in diseases with asymptomatic phases is particularly challenging and can take a long time, as their validation is by necessity often linked to long-term clinical outcomes. **[3a, 3b, 3c]**

30. Glassman RH, Ratain MJ. Biomarkers in early cancer drug development: limited utility. Clin Pharmacol Ther. 2009;85:134-5.

The authors of this short think piece offer a healthy skepticism of the utility of biomarkers in current drug development research. They suggest that the goals of early clinical trials are to: determine whether the formulation is acceptable; screen for toxicities not predicted by preclinical toxicology studies; assess relationships among dose, schedule, plasma concentrations, toxicities, and efficacy; and determine whether an expensive phase III program should be initiated. They argue that thus far, biomarker studies are neither necessary nor sufficient to meet these needs. A positive effect on a biomarker is not sufficient to conclude that a drug has antitumor activity or that a dose is optimal. Although biomarkers are potentially important tools in early clinical trials in oncology, to date they have not been shown to make cancer drug development more efficient or effective. The authors conclude that biomarker use should be limited to situations in which there is a strong scientific basis, a testable hypothesis, and a valid, precise assay method. **[1a, 3a, 3b]**

31. Gollob JA, Bonomi P. Historic evidence and future directions in clinical trial therapy of solid tumors. Oncology (Williston Park). 2006;20:10-8.

This paper provides an overview of treatments in renal cell cancer (RCC), melanoma and various solid tumors. The authors explain that the FDA has accepted significant results in clinical trials using surrogate endpoints as the basis for drug approval including the amount of tumor reduction, or tumor response (TR). They argue that although TR would seem to be a necessary precondition for improved survival, clinical studies have not consistently demonstrated a correlation between the two in patients with RCC. Moreover, TR may not be an appropriate endpoint for evaluating the effects of the new targeted therapies, whose putative mechanisms are generally cytostatic rather than cytotoxic. Clinical trials suggest that some patients with other solid tumors, such as lung cancer, may derive clinical benefit from treatment that helps stabilize their disease. The authors conclude that use of a variety of endpoints as well as different trial designs may provide an adequate basis for investigating the benefits/risks of newer therapies. **[3b, 3c]**

32. Goodsaid F, Frueh F. Biomarker qualification pilot process at the U.S. Food and Drug Administration. AAPS J. 2007;9:E105-8.
33. Goodsaid F, Papaluca M. Evolution of biomarker qualification at the health authorities. Nat Biotechnol. 2010;28:441-3.
34. Goodsaid FM, Frueh FW, Mattes W. Strategic paths for biomarker qualification. Toxicology. 2008;245:219-23.

The above three papers describe recent FDA initiatives concerning the qualification and regulatory evaluation of biomarkers. The authors draw a distinction between context-independent qualification (where proposed test methods can be qualified across multiple contexts of use) and context-dependent qualification processes. The authors suggest that while context-independence may be useful for research on toxicity biomarkers (especially for animal experiments), a context-dependent approach is required for biomarker qualification in drug development. For example, evidentiary standards for assessing preclinical and clinical data may vary according to whether a biomarker is predictive, diagnostic, or mechanistic. They describe a pilot process undertaken at the FDA consisting of a Biomarker Qualification Review Team that evaluated study protocols and reviewed study results for the qualification of novel biomarkers of drug safety, using appropriate preclinical, clinical, and statistical considerations. **[1a, 2a, 3a, 3b]**

35. Goulart BH, Clark JW, Pien HH, et al. Trends in the use and role of biomarkers in phase I oncology trials. Clin Cancer Res. 2007;13:6719-26.

This paper systematically reviewed abstracts submitted to the American Society of Clinical Oncology annual meeting (1991-2002) and the publications related to these abstracts to assess the use and role of biomarkers in phase I oncology trials. Twenty percent of American Society of Clinical Oncology phase I abstracts from 1991 to 2002 included biomarkers. This proportion increased over time (14 percent in 1991 compared with 26 percent in 2002). Independent predictors of the use of biomarkers included NCI sponsorship, submission in the time period of 1999 to 2002, adult population, and drug family (biological agents). Biomarkers supported dose selection for phase I studies in 11 of 87 of the trials (13 percent) emanating from these abstracts. Biomarker studies provided evidence supporting the proposed mechanism of action in 34 of 87 of the published trials (39 percent). The authors found that biomarkers made a minimal contribution to dose and schedule selection for phase I studies, but had greater effect in providing evidence confirming target modulation in human subjects. Their results suggest that acceptable toxicity and some evidence for antitumor effect remain the main end points used for decisions to proceed or not proceed with further drug development, and that use of biomarkers in Phase I studies is not warranted at this time. **[1a, 3a]**

36. Hampel H, Mitchell A, Blennow K, et al. Core biological marker candidates of Alzheimer's disease—perspectives for diagnosis, prediction of outcome and reflection of biological activity. J Neural Transm. 2004;111:247-72.

This review examines three candidate cerebrospinal fluids biomarkers for Alzheimer's disease (AD). The authors state that part of the rationale for the selection of these biomarkers was their ability to provide reasonable evidence for association with key mechanisms of pathogenesis or neurodegeneration in AD. At the time of their review, however, the preliminary and retrospective nature of the majority of findings, the absence of assay standardization, and the lack of comparison patient populations provided weak evidence for the usefulness of the biomarkers, particularly for predictive, diagnostic, or treatment evaluation purposes. **[1a, 2b, 3a]**

37. Hong H, Goodsaid F, Shi L, et al. Molecular biomarkers: a US FDA effort. Biomark Med. 2010;4:215-25.

This paper summarizes the current status of molecular biomarkers used for FDA-approved drug products, and discusses the challenges and future perspectives for the identification and qualification of molecular biomarkers. Specific FDA programs and research projects related to molecular biomarkers are also discussed for supporting regulatory review in the future. **[1a, 3b]**

38. Hunter DJ, Losina E, Guermazi A, et al. A pathway and approach to biomarker validation and qualification for osteoarthritis clinical trials. Curr Drug Targets. 2010;11:536-45.

This review outlines work done in other fields with regards biomarker validation and qualification and the lessons that may be learned by osteoarthritis (OA) researchers. The authors contend that defining a universally agreed upon path for biomarker validation and qualification is needed to address many of the challenges faced in OA drug development. These challenges include OA heterogeneity (patterns of onset and clinical presentation, different patterns of joint involvement, variations in rate of disease progression), and the current lack of a clear and reliably consistent disease modifying therapy. The paper proposes a qualification path that may be suitable for OA and presents concrete steps that might help achieve this. **[1a, 2b, 3a]**

39. Hurko O. The uses of biomarkers in drug development. Ann N Y Acad Sci. 2009;1180:1-10.

The paper, which focuses on diseases of the brain, suggests that although the value of surrogate markers is significant, such biomarkers are rare. Other, nonsurrogate biomarkers are, however, increasingly being used to reduce the risks of drug development. The author contends that any given biomarker is typically useful for only one of four types of risk reduction: (1) an inappropriate dosing regimen; (2) enrollment of nonresponsive subjects into clinical trials; (3) an inability to detect an efficacy signal quickly and reliably in chronic disorders; or (4) delayed recognition of potential side effects and/or toxicity. The author argues that a biomarker suitable for one purpose is usually not suitable for the other three. Although these considerations apply to all drug development, both the need and availability of appropriate biomarkers in each category vary between therapeutic areas. **[3c, 6]**

40. Jacobs A. An FDA perspective on the nonclinical use of the X-Omics technologies and the safety of new drugs. Toxicol Lett. 2009;186:32-5.

This paper explores the use of “omics” platforms for the evaluation of general toxicology, reproductive toxicology, the carcinogenicity potential of pharmaceuticals. The authors contend that though significant progress has been made in the standardization of procedures, challenges remain for evaluation of pharmaceuticals for regulatory purposes, because of off-target toxicologic effects, as well as issues of interpretation and the large number of biologic variables that can affect results (species/strain, genetic variations, diet, age, dose, duration, and weight of animals). They argue that variables also confound database compilations of expression profiles, and that the most promising use in the near future would be to clarify pathways for the various types of toxicity and carcinogenicity and get biomarkers for these pathways, to help assess relevance of nonclinical findings to humans. **[1a, 3a]**

41. Jain KK. Cancer biomarkers: current issues and future directions. Curr Opin Mol Ther. 2007;9:563-71.

This paper reviews cancer biomarkers and their role in understanding the pathobiological mechanisms of cancer as well as providing targets for drug discovery. It also examines the characteristics of an ideal cancer biomarker (summarized in Table 2) and emerging technologies for biomarker detection. The authors particularly focus on the use of biomarkers for anticancer drug development and clinical applications, including determination of prognosis as well as monitoring of response to therapy. The authors suggest that a major challenge in development of cancer biomarkers is that a number of genes are up- and down-regulated in cancer, making it problematic to rely on any single tumor biomarker even for one type of cancer. They argue that the physiological properties of the microenvironment of a majority (90 percent) of tumors, such as hypoxia, acidity, and changes in temperature, are considered promising environmental markers for tumor targeting. **[2a, 2b, 3b]**

42. Katz R. Biomarkers and surrogate markers: an FDA perspective. NeuroRx. 2004;1:189-195.

This article discusses the regulatory context and epistemological problems related to the interpretation of clinical trials in which unvalidated surrogate markers are used as primary outcomes. While the current law and regulations permit the FDA to base the approval of a drug product on a determination the effect of the drug on an unvalidated surrogate marker, there are a number of difficulties in interpreting trials that use surrogate markers as primary measures of drug effect. The author argues that because our knowledge of the relevant pharmacologic and biologic events is always imperfect/incomplete, drugs are typically approved on the simple finding of a beneficial effect in clinical trials (and on adequate safety data). This is ideal, he argues, not only because it is a direct clinical benefit that is obviously desired by the patient, but also because waiting to fully understand the relevant biologic events before drugs are approved is an undesirable strategy. This is why approval of a drug on the basis of an effect on an unvalidated surrogate marker represents such a fundamental departure from the typical course of action in drug approval. That is, approval of a drug on the basis of such an effect presupposes knowledge of events that is normally not only absent, but, in a sense, irrelevant. Because we do not ever have all of this knowledge, approval based on effects on surrogate markers will invariably involve a level of uncertainty not typical of the more standard route to drug approval. **[1a, 3a, 3b, 3c]**

43. Kelloff GJ, Sigman CC. New science-based endpoints to accelerate oncology drug development. Eur J Cancer. 2005;41:491-501.

This paper presents definitions and classifications of biomarkers for use in oncology drug development. Recent progress along with advances in imaging and bioassay technologies are the basis for describing and evaluating new biomarker endpoints as well as for defining other biomarkers for identifying patient populations, potential toxicity, and providing evidence of drug effect and efficacy. Science-based and practical criteria for validating biomarkers have been developed including considerations of mechanistic plausibility, available methods and technology, and clinical feasibility. For BMEBM purposes, these criteria for mechanistic plausibility—summarized in Table 3—are most relevant. **[1a, 3a, 3b]**

44. Kluft C. Principles of use of surrogate markers and endpoints. Maturitas. 2004;47:293-8.

The author of this paper—concerned with the validation of candidate surrogate end-points for the cardiovascular risk of sex steroids — contends that surrogate end-points are markers of biological mechanisms and require a mechanistic view on diseases. The validation process for venous thromboembolic disease and arterial disease biomarkers requires that a separate evaluation be made of the multiple clinical endpoints in which different biological mechanisms are likely to operate. Sex steroids have many effects on biological mechanisms and the selection and validation of surrogates from the changes in multiple mechanisms is a large enterprise. **[1a, 3c]**

45. Krishna R, Herman G, Wagner JA. Accelerating drug development using biomarkers: a case study with sitagliptin, a novel DPP4 inhibitor for type 2 diabetes. AAPS J. 2008;10:401-9.

In this review, applications of proximal (target engagement) and distal (disease-related) biomarkers are highlighted using the example of the recent development of sitagliptin for type 2 diabetes. The authors suggest that elucidation of target engagement and disease-related biomarkers significantly accelerated sitagliptin drug development, and facilitated design of clinical trials while streamlining dose focus and optimization, the net impact of which reduced overall cycle time to filing as compared to the industry average. [**1a, 3c]**

46. Kuhlmann J, Wensing G. The applications of biomarkers in early clinical drug development to improve decision-making processes. Curr Clin Pharmacol. 2006;1:185-91.

In this paper, the author contends that biomarkers are most useful in the early phase of clinical development when measurement of clinical endpoints or true surrogates may be too time-consuming or cumbersome to provide timely proof of principle or dose-ranging information. The use of biomarkers in early drug development helps to streamline clinical development by determining whether the drug is reaching and affecting the molecular target in humans, delivering findings that are comparable to preclinical data, and by providing a measurable endpoint that predicts desired or undesired clinical effects. Critical decisions such as candidate selection, early proof of mechanism or proof of concept, dose ranging and patient stratification as well as the assessment of development risks regarding safety, toxicity and drug interactions can be based on measurement of appropriate biomarkers that are biologically and/or clinically validated. Preclinical and phase I development plans can be focused to support an early biomarker study in healthy volunteers or mildly diseased patients, thus saving both resources and time. Dose estimates and patient stratification may reduce the size and duration of clinical studies in later phases of development, and safety and toxicity biomarkers may help to stop or continue a program early on. Even if a biomarker fails in the validation process there may still be a benefit of having used it as more knowledge about pathophysiology of the disease and the drug may be obtained. [**1a, 3a, 3b, 3c, 6]**

47. Kumar S, Mohan A, Guleria R. Biomarkers in cancer screening, research and detection: present and future: a review. Biomarkers. 2006;11:385-405.

This review describes the development of biomarkers in cancer research and detection with emphasis on different proteomic tools for the identification and discovery of new biomarkers, different clinical assays to detect various biomarkers in different specimens, role of biomarkers in cancer screening and the challenges in this direction of cancer research. Biomarkers offer a means for homogeneous classification of a disease and risk factor, and can extend basic information about the underlying pathogenesis of disease. The goals in cancer research include finding biomarkers that can be used for the early detection of cancers, design individual therapies, and to identify underlying processes involved in the disease. The author suggests that because so many myriad processes are involved in the diseased states, the goal is similar to finding a needle in a haystack. However, the development of ‘omic technologies, has allowed us to monitor a large number of key cellular pathways simultaneously. This has enabled the identification of biomarkers and signaling molecules associated with cell growth, cell death and cellular metabolism, and are also facilitating in monitoring the functional disturbance, molecular and cellular damage, and damage response. Table 3 (characteristics of an ideal screening program) and Table 5 (characteristics of an ideal biomarker) are particularly relevant for BMEBM purposes. **[1a, 3a, 3c]**

48. Kurian S, Grigoryev Y, Head S, et al. Applying genomics to organ transplantation medicine in both discovery and validation of biomarkers. Int Immunopharmacol. 2007;7:1948-60.

This paper examines the development of biomarkers in the area of organ transplantation. The authors state that the life and death nature of end stage organ failure, the severe donor organ shortage, and the powerful and toxic drug therapies required for the lifetimes of transplant patients, require biomarkers as tools to diagnose disease in its early stages, predict prognosis, suggest treatment options and then assist in the implementation of therapies. The article usefully highlights 2 broad approaches to biomarker discovery: (1) a hypothesis-driven approach in which a candidate molecule provides the target for which a biomarker is developed, and (2) looking for combinations or relatively large panels of biomarkers to diagnose disease conditions. The authors largely eschew the former approach for mechanistic reasons (e.g., complex interplay of many pathways, cascades and networks of molecules; cells or tissues interactions with neighboring cells that can trigger various cellular responses at a distant site). The authors argue that a single molecule or several in a single network cannot possibly direct and regulate these complex responses. Interestingly, they endorse the latter approach, which explores complex molecular networks and significantly increases the dimensionality of biomarker discovery using genomic technologies. This is an agnostic approach that eliminates the need to identify targets based a priori on specific biological knowledge of mechanisms and associations. [**1a, 1b, 2a, 2b, 3a]**

49. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. Eur J Cancer. 2007;43:2559-79.

This study reviewed a meta-analysis and individual articles concerning cancer prognostic biomarkers and found that published articles almost ubiquitously highlighted significant prognostic associations. In the rare articles where no prognostic markers were presented as significant, authors often presented other (nonprognostic) statistically significant analyses, expanded on the importance of nonsignificant trends, or defended the importance of the cancer marker with other arguments. Entirely ‘negative’ articles on prognostic cancer markers represented less than 1.5 percent of this literature. The authors conclude that under strong reporting bias, statistical significance loses its discriminating ability for the importance of prognostic markers. **[1a]**

50. Lassere MN. The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. Stat Methods Med Res. 2008;17:303-40.
51. Lassere MN, Johnson KR, Boers M, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. J Rheumatol. 2007;34:607-15.

In the above two articles, Lassere and collegues report on their efforts to develop a levels of evidence schema to evaluate biomarker and surrogate research. In the 2007 article, Lassere et al. propose a “variables in medicine” continuum. The lowest end is a large pool of biomarkers which are indicators of biological, pathological, or genetic disease development, or are a reflection of the mechanism of therapeutic interventions. As one moves forward on the continuum, the “variables container” becomes smaller when one considers only biomarkers that are validated clinical correlates or risk/prognostic factors. Then the “variables container” is narrowed further for the collection of surrogate endpoints. The highest end of the continuum consists only of very few outcomes which can be deemed as clinical endpoints in clinical trials. These endpoints reflect how a patient feels, functions, or survives. The constriction process of the “medical variable pool” is guided by the strength of the linkage of the biomarkers to the clinical endpoint, and by how confidently the biomarker can represent the clinical endpoint. The validation of risk/prognostic factors requires the demonstration of strong patient-level correlation only between the biomarker and the clinical endpoint.

This 2008 article represents a minor refinement of the above approach, now named the Biomarker-Surrogacy Evaluation Schema. The schema incorporates the three independent domains: Study Design, Target Outcome and Statistical Evaluation. Each domain has items ranked from0 to 5. The total score (0-15) determines the level of evidence, with Level 1 the strongest and Level 5 the weakest. The term “surrogate” is restricted to markers attaining Levels 1 or 2 only. Surrogacy status of markers can then be directly compared within and across different areas of medicine to guide individual, trial-based or drug-development decisions. **[1a, 3a, 3c]**

52. Lathia CD, Amakye D, Dai W, et al. The value, qualification, and regulatory use of surrogate end points in drug development. Clin Pharmacol Ther. 2009;86:32-43.

This article reviews the history of the development, qualification, and acceptance of nine surrogate endpoints. The authors argue that both the successes and failures had three key characteristics: (i) apparent biologic plausibility, (ii) prognostic value for the outcome of the disease, and (iii) an association between changes in the surrogates and changes in outcome with therapeutic intervention. With regard to biologic plausibility, the authors concluded that there was not sufficient evidence to make plausibility an absolute requirement for surrogate status, as the strength of plausibility did not appear to discriminate between the successes and the failures. They further argued that: “There is also a downside to making plausibility a requirement: new molecular ‘signature’ biomarkers that are identified by complex mathematical and statistical methods will not necessarily have initial plausibility that can be understood in conventional terms. Although concerns about unknown plausibility with such ‘black-box’ markers could lead to an increased requirement for more prognostic and/or clinical outcome linkage data, we would not seek to block them from surrogate status or other advanced regulatory use merely because we cannot mechanistically explain why they work, as long as their utility is tested in practice.” This article may be an especially useful touchstone for BMEBM thinking re: biomarkers and surrogates. **[1a, 3a, 3c]**

53. Lavallie ER, Dorner AJ, Burczynski ME. Use of ex vivo systems for biomarker discovery. Curr Opin Pharmacol. 2008;8:647-53.

This review article addresses the uses of ex vivo systems for both disease tissues and surrogate normal tissues to provide mechanistic insights into drug action and for the purpose of identifying candidate biomarkers. The authors make the case that biomarkers that either indicate PD effects or constitute predictive measures of individual patient responses can support dose selection and/or help determine therapeutic options. The development of biomarkers for clinical testing and validation can be facilitated by the use of ex vivo systems utilizing clinically relevant human tissues for the discovery of biomarkers of drug activity before first in human studies. Ex vivo analyses of tumors, liver tissue or hepatocytes, skin, and chondrocytes are discussed. **[2b]**

54. Lee JW, Figeys D, Vasilescu J. Biomarker assay translation from discovery to clinical studies in cancer drug development: quantification of emerging protein biomarkers. Adv Cancer Res. 2007;96:269-98.

This chapter discusses the challenges faced during cancer biomarker discovery as well as during technology and process translation, including pre-analytical planning, assay development, and preclinical and clinical validation. Particularly useful for BMEBM purposes is Figure 1, which describes intertwined processes of drug development and biomarker development. The horizontal blocks in the figure depict the progression of drug development of a new chemical or biological entity with unconfirmed mechanism of action. The drug development uses multiple biomarkers in various purposes from efficacy/safety assessment, down to market differentiation. The vertical blocks depict the developmental processes of moving a novel biomarker of unconfirmed mechanism to proof of biology, and to surrogacy. The processes include biomarker selection of on- and off-target markers, method development, validation, and application. **[1a, 3a, 3b, 3c]**

55. Lesko LJ. Paving the critical path: how can clinical pharmacology help achieve the vision? Clin Pharmacol Ther. 2007;81:170-7.

This article discusses the FDA’s Critical Path Initiative (CPI) and the role of clinical pharmacologists in the drug development process. It includes a discussion of model-based, semi-mechanistic drug development, drug/disease models that facilitate informed clinical trial designs and optimal dosing, the qualification process and criteria for new biomarkers and surrogate endpoints. Interestingly, the author argues that the CPI requires a paradigm shift from empiric to mechanistic thinking to improve the efficiency, predictiveness, and informativeness of clinical drug development. This shift, he argues, is driven by more precise biological and molecular definitions of disease phenotypes where traditional diseases are being subdivided into different subtypes. This, in turn, can lead to more targeted treatments with significantly improved benefit/risk ratios. Figures 2 and 3 in the paper—flowcharts describing key questions and qualification processes for biomarkers—may be useful for BMEBM purposes. **[1a, 3a, 3b]**

56. Lock EA, Bonventre JV. Biomarkers in translation; past, present and future. Toxicology. 2008;245:163-6.

This paper discusses the history of the discovery of biomarkers for renal and cardiac injury. The authors summarize the use of biomarkers in preclinical evaluation in experimental animals and in patients to help diagnose or monitor a disease, predict outcome or to evaluate a therapeutic intervention. **[2a, 3b]**

57. Lonn E. The use of surrogate endpoints in clinical trials: focus on clinical trials in cardiovascular diseases. Pharmacoepidemiol Drug Saf. 2001;10:497-508.

This review provides a definition of surrogate endpoints, proposes practical criteria for establishing their validity, outlines some of the advantages, disadvantages and specific statistical considerations associated with their use in clinical trials and attempts to also highlight drug approval issues associated with the use of these endpoints. A number of examples are also provided related to the use of surrogate endpoints in clinical trials with special emphasis on their use in cardiovascular medicine. **[1a, 3c]**

58. Marrer E, Dieterle F. Biomarkers in oncology drug development: rescuers or troublemakers? Expert Opin Drug Metab Toxicol. 2008;4:1391-402.

59. Marrer E, Dieterle F. Promises of biomarkers in drug development—a reality check. Chem Biol Drug Des. 2007;69:381-94.

The above two articles discuss different types of biomarkers, their identification, validation and use in different phases of drug development from drug discovery, to approval, to clinical application, as well as the state-of-the-art biomarker technologies and promising future methods. Though they do not offer great insight into key questions raised by BMEBM, they provide a useful overview of the state of the field, and address some economic considerations that that affect biomarker development and implementation. **[1a]**

60. McMichael AJ, Hall AJ. The use of biological markers as predictive early-outcome measures in epidemiological research. IARC Sci Publ. 1997;281-9.

This article examines the uses of biomarkers in epidemiological research as early-outcome measures to predict the occurrence of clinical disease and to elucidate the biological mechanism of pathogenesis. The authors suggest that the proposed epidemiological use is conceptually less straightforward than the well-established use of biomarkers to improve or extend exposure assessment or to study inter-individual variations in disease susceptibility. In principle, they argue that this form of use could accelerate or facilitate etiological research. The authors suggest that this mode of biomarker use, especially in cancer epidemiology, is the least clear-cut and the least well developed. The recurrent problem is identifying biomarkers that: (1) are on the causal pathway, (2) have a high probability of progression to clinical disease, and (3) account for all or most of the cases of the specified clinical outcome. Such biomarkers would be most useful if they conferred a long lead-time relative to clinical disease occurrence. **[1a, 3a, 3b]**

61. Merlo DF, Sormani MP, Bruzzi P. Molecular epidemiology: new rules for new tools? Mutat Res. 2006;600:3-11.

The authors of this article begin by defining molecular epidemiology as a field that "combines biological markers and epidemiological observations in the study of the environmental and genetic determinants of cancer and other diseases." They argue that there are many advantages associated with incorporating biomarkers into epidemiologic research, including: (1) increased sensitivity and specificity to carcinogenic exposures; (2) more precise evaluation of the interplay between genetic and environmental determinants of cancer; (3) earlier detection of carcinogenic effects of exposure; (4) characterization of disease subtypes-etiologies patterns; (5) evaluation of primary prevention measures. They suggest that an area that has not received sufficient attention concerns the validation of these biomarkers as surrogate endpoints for cancer risk. Of particular relevance to BMEBM is the authors' contention that the challenges posed by the application of validation principles to epidemiological research, where the basic tool for this validation (i.e., the randomized study) is seldom possible, have not been thoroughly explored. They examine a number of observational study designs that may help to address this challenge. This article provides a good companion piece to McMichael and Hall (1997). **[1a, 3a, 3b]**

62. Oldenhuis CN, Oosting SF, Gietema JA, et al. Prognostic versus predictive value of biomarkers in oncology. Eur J Cancer. 2008;44:946-53.

This article seeks to clarify differences between prognostic and predictive cancer biomarkers. The authors contend that the best prognostic factors are still simple clinical parameters (e.g., performance status, number of metastatic sites, tumor grade), and that prognostic biomarkers might be useful for hypothesis testing for their relevance as predictive markers, as targets for therapy and for the selection of patients for adjuvant treatment. A predictive factor is used upfront to predict response to therapy or is monitored during treatment to define the effectiveness of this treatment. They argue that predictive biomarkers are needed that can guide patient tailored therapy as knowledge of biological mechanisms of tumors and evidence of their heterogeneity and multi-factorial nature grows. Most studies only contribute low levels of evidence due to retrospective data and small sample size, and many reports lack sufficient information to be compared to other studies. The authors believe that there is an urgent need for prospective data to validate hypotheses and it is therefore of great importance that biomarker analyses are incorporated into randomized clinical trials as a separate objective. **[1a, 3a]**

63. Park JW, Kerbel RS, Kelloff GJ, et al. Rationale for biomarkers and surrogate end points in mechanism-driven oncology drug development. Clin Cancer Res. 2004;10:3885-96.

This article highlights the potential, as well as the associated challenges, of using mechanism-based biomarkers to facilitate the development of molecular targeted therapies. The authors contend that targeting drugs to molecularly defined populations is difficult to implement and has not been a traditional approach in the development of new drugs. By providing insight into disease mechanisms and interactions with therapy, the successful implementation of biomarkers can significantly advance the effort to rationally develop targeted agents. Case studies of trastuzumab, imatinib, EGFR inhibitors and angiogenesis inhibitors are examined. **[3a, 3c]**

64. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst. 2001;93:1054-61.

The purpose of this article is to define a formal structure to guide the process of biomarker development. The authors categorize the development into five phases that a biomarker needs to pass through to produce a useful population-screening tool. The phases of research are generally ordered according to the strength of evidence that each provides in favor of the biomarker, from weakest to strongest. The five phases are: Phase 1—Preclinical Exploratory Studies; Phase 2—Clinical Assay and Validation; Phase 3—Retrospective Longitudinal Repository Studies; Phase 4—Prospective Screening Studies; and Phase 5—Cancer Control Studies. As the authors note, evidential criteria for when a biomarker can reasonably progress from one phase of development to the next was not included in their proposal, and require multidisciplinary panels of experts for their definition. The authors make clear that their proposal is not meant to be rigid or definitive; rather it a foundation for dialogue that will ultimately lead to improved rigor in biomarker research. **[1a, 3a]**

65. Psaty BM, Lumley T. Surrogate end points and FDA approval: a tale of 2 lipid-altering drugs. JAMA. 2008;299:1474-6.
66. Psaty BM, Weiss NS, Furberg CD, et al. Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. JAMA. 1999;282:786-90.

The above two articles examine surrogate end point approaches to the drug approval process in the treatment of cardiovascular disease. At issue is the need to balance the public health advantages of rapid approval for drugs that turn out to be safe and effective against harms that might occur when drugs approved on the basis of surrogate end points turn out later either to have significant safety problems or to lack efficacy.

The 1999 article examines the advantages and disadvantages of using surrogates to assess therapies. The authors state that reliance on surrogates raises the possibility of incomplete, inadequate, or misleading evaluations, stating: "To use only a surrogate end point is to accept as empirical evidence for clinical practice a hypothesis about health benefits that has never been tested." They offer two proposals for reforming the drug approval process.

The 2008 article examines recent experience with two lipid-altering drugs, ezetimibe and torcetrapib. The authors argue that the rapid FDA approval and aggressive marketing of ezetimibe before RCTs were underway exposed patients to unnecessary risks. By contrast, an RCT for torcetrapib was underway before FDA approval, and the trial was stopped early because of an increase in the risk of the primary end point, major cardiovascular events. **[1a, 3c]**

67. Ransohoff DF. Evaluating discovery-based research: when biologic reasoning cannot work. Gastroenterology. 2004a;127:1028.

This brief editorial makes the emphatic point that, in the area of the new 'omics technologies, questions about biological plausibility — namely questions in the form of "Should it work?"—are off-target. Rather, the question about whether something does or does not work can be settled– and must be settled– by a different kind of reasoning. This form of reasoning requires direct observation and comparison by using research methods that rigorously avoid chance and bias as alternate explanations for results. The author argues that problems of chance and bias are primarily the domain not of biologic reasoning but of epidemiologic reasoning. **[1a]**

68. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer. 2004b;4:309-14.
69. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer. 2005;5:142-9.
70. Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. J Clin Epidemiol. 2007;60:1205-19.
71. Ransohoff DF. Promises and limitations of biomarkers. Recent Results Cancer Res. 2009;181:55-9.

The above four articles examine the promise and limitations of the 'omics technologies in the area of cancer biomarker development. The author draws an interesting distinction between drug discovery and biomarker discovery research. Drug research typically proceeds through prospective randomized and blinded studies that make comparisons between treated and nontreated subjects, and so, in the author's words, "tends to involve generally reliable studies." Biomarker studies (for markers of prognosis or diagnosis) are observational, and "are routinely threatened by serious bias." As such, the author repeatedly beats the drum in support of investigators learning to understand the sources of bias affecting biomarker studies and how to approach them. Researchers often rely on tools such as “phases” and “guidelines” to facilitate communication and collaboration (see Pepe 2001), believing that they provide thorough prescriptions for conducting clinical research. The author believes that these tools are limited, and that much greater attention should be paid to study design (the details that constitute the substance of the actual planning, conduct, and interpretation of a research study) of each individual study. [**1a, 3b]**

72. Ratain MJ, Glassman RH. Biomarkers in phase I oncology trials: signal, noise, or expensive distraction? Clin Cancer Res. 2007;13:6545-8.

This is an editorial commentary in response to the meta-analysis by Goulart et al. (2007). Although the authors take issue with the operational definition of biomarker Goulart et al. use and raise other limitations of their study, they accept Goulart at al.’s conclusions that the use of biomarkers in Phase I oncology studies is not warranted at this time. **[1a]**

73. Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. Int J Clin Oncol. 2009;14:102-11.

This article addresses general issues and challenges surrounding surrogate endpoints in cancer clinical trials. The authors review various meta-analytic methodologies concerning the application of these methods to cancer clinical trials with different tumor types. In oncology, several applications have successfully identified useful surrogates (e.g., disease-free survival and progression-free survival as surrogates for overall survival in advanced colorectal cancer). They also discuss several limitations of surrogate endpoints, including issues related to the extrapolation of the validity of a surrogate. They suggest that the success of the surrogate is very likely to depend on specific individual mechanisms of the treatment, and caution that the extrapolation of the validity of one surrogate into different disease populations, interventions with different biological pathways, and sometimes even only different dose levels, may not be appropriate. The authors contend that before applying a surrogate endpoint in a trial, the benefit-risk tradeoff should always be examined thoroughly, based on historical clinical trial results and new evidence available**. [1a, 3c]**

74. Sinha A, Singh C, Parmar D, et al. Proteomics in clinical interventions: achievements and limitations in biomarker development. Life Sci. 2007;80:1345-54.

This review article examines achievements and limitations of proteomics in developing predictive biomarkers for toxicological and clinical interventions. They authors contend that proteomics is providing insights into the mechanism of action of a wide range of substances and is being used to increase speed and sensitivity of toxicological screening by identifying toxicity and efficacy biomarkers. The major challenges, they argue, are the discrimination of changes due to inter-individual variation, experimental background noise in protein profiling, and post-translational modifications. Despite intensive research, only a very limited number of plasma proteins have been validated as biomarkers for disease. Although proteome approaches have provided opportunities to define molecular mechanisms of toxicity and clinical interventions, reproducibility in expression depends on experimental conditions across different laboratories and, therefore, remains a challenge for the field. **[1a, 3a]**

75. Sistare FD, Dieterle F, Troth S, et al. Towards consensus practices to qualify safety biomarkers for use in early drug development. Nat Biotechnol. 2010;28:446-54.

This article describes core principles, mutual decisions, and unresolved issues which emerged in the course of the Predictive Safety Testing Consortium's efforts to qualify seven new safety biomarkers of drug-induced renal injury to support regulatory decision-making during early drug development. They articulate a number of strength-of-evidence criteria for evaluating safety biomarkers, including: biological plausibility of the association of the biomarkers with injury to the organ of interest; understanding of the molecular mechanism of the biomarker response; strong association of changes in biomarker levels to pathological outcomes and superior performance relative to currently accepted biomarkers; and consistent response across mechanistically diverse toxicants, sexes, strains, and species (see Box 1). The article does a nice job in articulating the scientific and economic challenges of qualifying biomarkers, making the case for establishing a consortium of stakeholders from industry, academia and regulatory bodies. **[1a, 3a, 6]**

76. Tan DS, Thomas GV, Garrett MD, et al. Biomarker-driven early clinical trials in oncology: a paradigm shift in drug development. Cancer J. 2009;15:406-20.

This article provides an overview of biomarkers in early clinical trials, including examples where they have been particularly successful, and the caveats and pitfalls associated with indiscriminate application. The authors describe the use of PD endpoints to demonstrate the proof of modulation of target, pathway, and biologic effect, as well as predictive biomarkers for patient selection and trial enrichment. They contend that accurate preclinical models are important for PK-PD-efficacy modeling and biomarker validation. The degree of scientific and analytical validation should ensure that biomarkers are fit-for purpose, according to the stage of development and the impact on the trial; specifically they are either exploratory or used to make decisions within the trial. To be maximally useful at an early stage, these must be in place before the commencement of phase I trials. Validation and qualification of biomarkers then continues through clinical development. The authors highlight the impact of technology platforms such as genomics, proteomics, circulating tumor cells, and minimally invasive functional and molecular imaging, with respect to their potential role in improving the success rate and speed of drug development and in interrogating the consequences of therapeutic intervention and providing unique insights into human disease biology. **[1a, 2a, 3a]**

77. Taube SE, Clark GM, Dancey JE, et al. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. J Natl Cancer Inst. 2009;101:1453-63.

This article reports the issues discussed and recommendations issued from a 2007 workshop sponsored by the National Cancer Institute and the US Food and Drug Administration concerning past lessons learned and ongoing challenges faced in biomarker development and drug and biomarker codevelopment. Figure 1 of the paper presents a useful schematic of the considerations addressed, including evidence/understanding of biological mechanism. According to the authors, a recurring theme in the workshop's discussions was the need to increase understanding of the biology associated with the chosen biomarkers, including their role in tumor behavior and their interplay with the drugs ' mechanisms of action. This biological understanding should be integrated into the clinical context in which the biomarker would be used (see case studies of HER2 and EGFR summarized in Boxes 1 and 2, respectively). “Biological rationale” is described as a key factor in selecting candidate predictive biomarkers for codevelopment with a cancer therapeutic. This would include evidence that the biomarker(s) chosen is meaningfully correlated with the activity of the targeted agent. The evidence may derive from existing preclinical and clinical literature; data collected during phase I, expanded phase II, or early phase III clinical studies of the agent; or modeling and biological inferences using incomplete or partially complete provisional datasets and data mining. All available information on cellular pathways relevant to drug action, mechanisms, or drug interactions should be included in decision making and in generation of study hypotheses. The authors make clear that some of the major challenges for biomarker assay and therapeutic agent codevelopment relate to the costs of standardizing and evaluating the utility of an assay when neither the clinical activity of the agent nor the relationship of the biomarker to the mechanism(s) of action of the agent being developed is clear. **[1a, 2a, 2b, 3a]**

78. Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? JAMA. 1999;282:790-5.

In this article, the author makes the case that to become surrogates biomarkers need to be correlated with outcome in clinical trials of more than one drug with the same mechanism of action targeted at the same indication. See Table 1 for factors related to biological plausibility that either favor or do not favor candidate surrogates. **[1a, 3c]**

79. Trusheim MR, Berndt ER, Douglas FL. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. Nat Rev Drug Discov. 2007;6:287-93.

This article examines current and emerging examples in which therapies are matched with specific patient population characteristics using clinical biomarkers (which the authors call “stratified medicine”) and discusses the implications of this approach to future drug development strategies and market opportunities. The authors suggest that “differential biological mechanism” is a necessary condition for stratified medicine, claiming that at least one of the following biological characteristics with the potential to differentiate patients must exist: (1) underlying disease variability reflecting multi-factorial etiology, or currently indistinguishable clinical presentations for biologically distinct conditions; (2) multiple relevant targets for medical intervention; (3) differential ADME (absorption, distribution, metabolism and excretion) characteristics, toxicity or tolerability of the therapeutic regimen(s); and (4) adaptiveness of the disease leading to treatment resistance. **[1a, 3a, 3b]**

80. van Gool AJ, Henry B, Sprengers ED. From biomarker strategies to biomarker activities and back. Drug Discov Today. 2010;15:121-6.

This review outlines the rational question-based drug development strategy in which biomarker data drive decisions on which drug candidates to progress to clinical testing. The authors cite the high attrition rates during clinical development, arguing that the major cause for this high attrition is the proof-of-concept phase, during which the attrition rate is as high as 80 percent because of a lack of efficacy and/or unacceptable safety liabilities. This leads them to surmise that preclinical studies in pharmaceutical research are insufficient to predict drug action in patients. The authors then proceed to outline their “question-based” drug development approach, which includes a number of mechanistic considerations. One question included in their approach is: “Does the compound cause its intended pharmacological and functional effects?” Most drugs inhibit or stimulate their target, with downstream functional consequences. To verify that the drug has the same effect on the pathway in patients as in the preceding cellular or animal models, clinically applicable biomarkers indicative of signaling pathways are required. For example, for a kinase drug target active in blood cells, one can monitor the phosphorylated substrates and/or downstream gene and protein expression patterns. If the drug is found not to modulate its mechanism of action in patients, one can decide to switch to an improved drug candidate. If the drug does modulate the mechanism as predicted but there is no effect on disease parameters, the concept of targeting this drug target in these patients can be abandoned. **[1a, 3a, 3b]**

81. Verma M, Srivastava S. New cancer biomarkers deriving from NCI early detection research. Recent Results Cancer Res. 2003;163:72-84; discussion 264-6.
82. Verma M, Wright GLJ, Hanash SM, et al. Proteomic approaches within the NCI early detection research network for the discovery and identification of cancer biomarkers. Ann N Y Acad Sci. 2001;945:103-15.

The above two articles discuss the efforts of the Early Detection Research Network at the National Cancer Institute, in bringing together scientific expertise from leading national and international institutions, to identify and validate biomarkers for the detection of precancerous and cancerous cells in determining risk for developing cancer. Other topics covered include the use of genomics and proteomics as high-throughput technology platforms to facilitate biomarker-aided detection of early cancer, and issues surrounding the analysis, validation, and predictive value of biomarkers using such technologies. **[1a, 2b, 3a]**

83. Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev. 2007;16:1954-65.
84. Vineis P, Porta M. Causal thinking, biomarkers, and mechanisms of carcinogenesis. J Clin Epidemiol. 1996;49:951-6.

The above two articles explore how the use of biomarkers should be considered within the context of causal models in epidemiology, and of the intertwining of causation and pathogenesis. The authors indicate that the use of biomarkers is increasing both in acute and chronic disease epidemiology, but the rationale for their introduction is not always firmly established. Unlike infectious diseases, for cancer and cardiovascular disease external "necessary" causes have not been identified. Thus, the classification of cancer and other chronic diseases cannot be based on unequivocal criteria such as the "etiologic" classification of infectious diseases. From a mechanistic point of view, unless molecular biology discovers specific mechanistic steps in carcinogenesis, which indicate the existence of "necessary" events in carcinogenesis, we cannot adopt an unequivocal definition of cancer. The authors argue that the potential contribution of biomarkers to the elucidation of the pathogenetic process should be considered in the light of such uncertainties. There is a range of indications for biomarkers, from the use of very specific measurements aimed at single molecules, to measurements indicating cumulative exposure to agents with the same mechanism of action. The potential uses of markers in chronic disease epidemiology include: exposure assessment in cases in which traditional epidemiologic tools are insufficient; multiple exposures or mixtures, in which the aim is to disentangle the etiologic role of single agents; estimation of the total burden of exposure to chemicals having the same mechanistic target; and investigation of pathogenetic mechanisms. **[1a, 3a, 3b]**

85. Wagner JA. Overview of biomarkers and surrogate endpoints in drug development. Dis Markers. 2002;18:41-6.
86. Wagner JA, Williams SA, Webster CJ. Biomarkers and surrogate end points for fit-for-purpose development and regulatory evaluation of new drugs. Clin Pharmacol Ther. 2007;81:104-7.
87. Wagner JA. Strategic approach to fit-for-purpose biomarkers in drug development. Annu Rev Pharmacol Toxicol. 2008;48:631-51.
88. Wagner JA. Biomarkers: principles, policies, and practice. Clin Pharmacol Ther. 2009;86:3-7.
89. Wagner JA, Prince M, Wright EC, et al. The Biomarkers Consortium: practice and pitfalls of open-source precompetitive collaboration. Clin Pharmacol Ther. 2010;87:539-42.

The above five articles provide an account of fit-for-purpose biomarker qualification. Wagner and colleagues suggest that thinking of qualification in terms of stages offers a useful way to understand the relationship between evidence evaluation and utility for decision-making.
On this account, biomarker qualification is a graded, fit-for-purpose evidentiary process linking a biomarker with biological processes and clinical endpoints, dependent on the intended application. Biomarkers used in drug development can be categorized into four classes of qualification: (a) Exploration biomarkers are research and development tools accompanied by in vitro and/or preclinical evidence, but with no consistent information linking the biomarker clinical outcomes in humans; (b) demonstration biomarkers are associated with adequate preclinical sensitivity and specificity and are linked with clinical outcomes, but have not been reproducibly demonstrated in clinical studies; (c) characterization biomarkers are associated with adequate preclinical sensitivity and specificity and are reproducibly linked to clinical outcomes in more than one prospective clinical study in humans; and (d) surrogacy reflects a holistic evaluation of the available data, demonstrating that the biomarker can substitute for a clinical endpoint. Observational data collected from clinical settings can also provide critical information in the qualification process, linking mechanistic data at the molecular level to population-level findings. This schema captures the increasing decision-making and regulatory utility of biomarkers as qualifying evidence is accrued. **[1a, 2a, 3a, 3b, 3c]**

90. Williams SA, Slavin DE, Wagner JA, et al. A cost-effectiveness approach to the qualification and acceptance of biomarkers. Nat Rev Drug Discov. 2006;5:897-902.

The authors of this article take the view that given the absence of widely accepted and practically applicable criteria that facilitate adequate biomarker qualification, cost-effectiveness considerations should be brought to bear on the evaluation of biomarkers. They assess existing qualification schemas and conclude that each is too subjective to achieve wide acceptance among a diverse set of stakeholders. They articulate a set of principles that enable cost-effectiveness evaluations of biomarkers even with incomplete knowledge (see Box 1). **[1a]**

91. Woodcock J. Chutes and ladders on the critical path: comparative effectiveness, product value, and the use of biomarkers in drug development. Clin Pharmacol Ther. 2009;86:12-4.
92. Woodcock J, Woosley R. The FDA critical path initiative and its influence on new drug development. Annu Rev Med. 2008;59:1-12.

The above two articles describe the FDA's Critical Path Initiative with respect to the role of biomarkers and biomarker qualification. Emphasis is placed on the need to foster public-private partnerships and consortia of experts in order to ensure that evaluation occurs in a timely and rigorous manner. [**1a]**

93. Zhao L, Jin W, Rader D, et al. A Translational Medicine perspective of the development of torcetrapib: Does the failure of torcetrapib development cast a shadow on future development of lipid modifying agents, HDL elevation strategies or CETP as a viable molecular target for atherosclerosis? A case study of the use of biomarkers and Translational Medicine in atherosclerosis drug discovery and development. Biochem Pharmacol. 2009;78:315-25.

This article uses a case study of the failed development of torcetrapib to emphasize the need for a paradigm shift from the conventional drug development mode to a biomarker-based Translational Medicine (TMed) strategy. Although the relationship between HDL (high density lipoprotein) function and cardiovascular (CV) risk had been extensively explored, the premise that HDL elevation is linked to reduced CV risks and that high HDL cholesterol (HDL-C) might be a potential surrogate biomarker for reduced CV risk remains controversial. Substantial genetic, molecular, biochemical and preclinical evidence raised the hope that HDL-C elevation via cholesteryl ester transfer protein (CETP) inhibition might generate clinical benefits. However, four large-scale clinical trials with the CETP inhibitor torcetrapib failed to demonstrate benefits on CV clinical outcomes. Likewise, biomarkers that were supposed to predict vascular risk reduction provided disappointing results. Emergence of further CETP inhibitors encourage continued development of such compounds for cardiovascular risk management. However, there is a need to adopt biomarker-driven TMed strategies in target validation, target-compound interaction, PD activities, disease modification and patient selection to guide future drug development efforts. These strategies may elucidate multiple, complex pathways and help yield more compelling mechanistic evidence for the relationship between HDL and prevention of CHD. **[1a, 2a, 2b, 3a, 3b]**

II. Articles Retrieved But Not Annotated

Category A: Added minimally to other information (i.e., limited or redundant)

1. Bast RCJ, Lilja H, Urban N, et al. Translational crossroads for biomarkers. Clin Cancer Res. 2005;11:6103-8.
2. Cavero I. Optimizing the preclinical/clinical interface: an Informa Life Sciences conference 12-13 December, 2006. Expert Opin Drug Saf. 2007;6:217-24.
3. Desai AA, Hysi P, Garcia JG. Integrating genomic and clinical medicine: searching for susceptibility genes in complex lung diseases. Transl Res. 2008;151:181-93.
4. Desar IM, van Herpen CM, van Laarhoven HW et al. Beyond RECIST: molecular and functional imaging techniques for evaluation of response to targeted therapy. Cancer Treat Rev. 2009;35:309-21.
5. Duffy SW, Treasure FP. Potential surrogate endpoints in cancer research—some considerations and examples. Pharm Stat. 2011;10(1):34-9
6. Fasolo A, Sessa C. Translational research in phase I trials. Clin Transl Oncol. 2009;11:580-8.
7. Levinson SS. Weak associations between prognostic biomarkers and disease in preliminary studies illustrates the breach between statistical significance and diagnostic discrimination. Clin Chim Acta. 2010; 411:467-73.
8. Lin D, Hollander Z, Meredith A, et al. Searching for ‘omic’ biomarkers. Can J Cardiol. 2009;25 Suppl A:9A-14A.
9. Lockhart BP, Walther B. [Biomarkers: “Found in translation”]. Med Sci (Paris). 2009;25:423-30.
10. Moore RE, Kirwan J, Doherty MK et al. Biomarker discovery in animal health and disease: the application of post-genomic technologies. Biomark Insights. 2007;2:185-96.
11. Nicolette CA, Miller GA. The identification of clinically relevant markers and therapeutic targets. Drug Discov Today. 2003;8:31-8.
12. Potter JD. Cancer prevention: epidemiology and experiment. Cancer Lett. 1997;114:7-9. .
13. Richter WS. Imaging biomarkers as surrogate endpoints for drug development. Eur J Nucl Med Mol Imaging. 2006;33 Suppl 1:6-10.
14. Sawyers CL. The cancer biomarker problem. Nature. 2008;452:548-52.
15. Steele VE, Kelloff GJ. Development of cancer chemopreventive drugs based on mechanistic approaches. Mutat Res. 2005;591:16-23.
16. Warnock DG, Peck CC. A roadmap for biomarker qualification. Nat Biotechnol. 2010;28:444-5.

Category B: Focused entirely on mechanisms of disease, with no attention to mechanisms of therapeutic interventions

Category C: Article could not be located

1. Gimmi CD. Current stumbling blocks in oncology drug development. Ernst Schering Res Found Workshop. 2007;135-49. **[C]**
2. Kelloff GJ, Boone CW, Crowell JA, et al. Surrogate endpoint biomarkers for phase II cancer chemoprevention trials. J Cell Biochem Suppl. 1994;19:1-9. **[C]**
3. Kensler TW, Davidson NE, Groopman JD, et al. Biomarkers and surrogacy: relevance to chemoprevention. IARC Sci Publ. 2001;154:27-47. **[C]**
4. Ryan NS, Fox NC. Imaging biomarkers in Alzheimer's disease. Ann N Y Acad Sci. 2009;1180:20-7. **[C]**
5. Strand KJ, Khalak H, Strovel JW, et al. Expression biomarkers for clinical efficacy and outcome prediction in cancer. Pharmacogenomics. 2006;7:105-15. **[C]**
6. Vainio H. Promise of molecular epidemiology—epidemiologic reasoning, biological rationale and risk assessment. Scand J Work Environ Health. 1999;25:498-504. **[C]**
7. Venitz J. Using exposure-response and biomarkers to streamline early drug development. Ernst Schering Res Found Workshop. 2007;47-63. **[C]**

III. Mapping articles onto conceptual framework

1. Strength of evidence for existence of intervention's pathway

c. Quality (design and execution) and strength (quantitative effect) of experimental evidence in preclinical models.

- Altar (2008a, 2008b)
- Alymani (2010)
- Fine (2009)
- Floyd (2004)
- Glassman (2009)
- Goodsaid (2007, 2008, 2010)
- Goulart (2007)
- Hampel (2004)
- Hong (2010)
- Hunter (2010)
- Jacobs (2009)
- Katz (2004)
- Kelloff (2005)
- Kluft (2004)
- Krishna (2008)
- Kuhlmann (2006)
- Kumar (2006)
- Kurian (2007)
- Kyzas (2007)
- Lassere (2007, 2008)
- Lathia (2009)
- Lee (2007)
- Lesko (2007)
- Lonn (2001)

- McMichael (1997)
- Merlo (2006)
- Oldenhuis (2008)
- Pepe (2001)
- Psaty (1999, 2008)
- Ransohoff (2004a, 2004b, 2005, 2007, 2009)
- Ratain (2007)
- Shi (2009)
- Sinha (2007)
- Sistare (2010)
- Tan (2009)
- Taube (2009)
- Temple (1999)
- Trusheim (2007)
- van Gool (2010)
- Verma (2001, 2003)
- Vineis (1996, 2007)
- Wagner (2002, 2007, 2008, 2009, 2010)
- Williams (2006)
- Woodcock (2008, 2009)
- Zhao (2009)

d. Number of experimental models

- Antoine (2009)
- Kurian (2007)

e. Variety of experimental models (e.g., animal species)

- Day (2008)

7. Strength of evidence that the pathway exists in human disease states.
   e. Strength of evidence for animal/in vitro model's relevance for human disease state.

- Boffetta (2010)
- Danhof (2005)
- Danna (2006)
- Day (2008)
- Fleming (2005)
- Floyd (2004)
- Goodsaid (2007, 2008, 2010)
- Jain (2007)
- Kurian (2007)
- Lock (2008)
- Tan (2009)
- Taube (2009)
- Wagner (2002, 2007, 2008, 2009, 2010)

- Zhao (2009)

f. Ex vivo evidence

- Bhattacharya (2009)
- Bhogal (2008)
- Chau (2008)
- Clark (2009)
- Coate (2009)
- Collins (2006)
- Danna (2006)
- Dhani (2008)
- Dudley (2009)
- Hampel (2004)
- Hong (2010)
- Hunter (2010)
- Jain (2007)
- Kurian (2007)
- Lavallie (2008)
- Taube (2009)
- Verma (2001, 2003)
- Wagner (2002, 2007, 2008, 2009, 2010)
- Zhao (2009)

g. Evidence that pathway occurs in complete physiologic system (e.g., functioning hearts vs. heart tissue.)

h. Evidence from human physiologic experiments.

8. Completeness of proposed mechanistic pathway. (From intervention to clinical endpoint)

a. Gaps in pathway (including whether intervention/exposure can exert effect on target due to issues of bioavailability, metabolism, delivery, etc.)

- Alymani (2010)
- Antoine (2009)
- Boffetta (2010)
- Carden (2009)
- Chau (2008)
- Chetty (2010)
- Coate (2009)
- Colburn (2000, 2003)
- Danna (2006)
- Day (2008, 2009)
- Fine (2009)
- Floyd (2004)
- Frank (2003)

- Glassman (2009)
- Goodsaid (2007, 2008, 2010)
- Goulart (2007)
- Hampel (2004)
- Hunter (2010)
- Jacobs (2009)
- Katz (2004)
- Kelloff (2005)
- Kuhlmann (2006)
- Kumar (2006)
- Kurian (2007)
- Lassere (2007, 2008)
- Lathia (2009)
- Lee (2007)
- Lesko (2007)
- McMichael (1997)
- Merlo (2006)
- Oldenhuis (2008)
- Park (2004)
- Pepe (2001)
- Sinha (2007)
- Sistare (2010)
- Tan (2009)
- Taube (2009)
- Trusheim (2007)
- van Gool (2010)
- Verma (2001, 2003)
- Vineis (1996, 2007)
- Wagner (2002, 2007, 2008, 2009, 2010)
- Zhao (2009)

b. Remoteness of the mechanistic outcomes from clinical outcomes.

- BDWG (2001)
- Carroll (2007)
- Clark (2009)
- Colburn (2000, 2003)
- Collins (2006)
- Day (2009)
- Dhani (2008)
- Frank (2003)
- Glassman (2009)
- Gollob (2006)
- Goodsaid (2007, 2008, 2010)
- Jain (2007)
- Katz (2004)
- Kelloff (2005)

- Kuhlmann (2006)
- Lee (2007)
- Lesko (2007)
- Lock (2008)
- McMichael (1997)
- Merlo (2006)
- Ransohoff (2004b, 2005, 2007, 2009)
- Trusheim (2007)
- van Gool (2010)
- Vineis (1996, 2007)
- Wagner (2002, 2007, 2008, 2009, 2010)
- Zhao (2009)

c. Strength of evidence linking proximal (i.e., surrogate) to distal (i.e., definitive) clinical endpoints

- Altar (2008a, 2008b)
- Alymani (2010)
- Bhattacharya (2009)
- Bhogal (2008)
- BDWG (2001)
- Boffetta (2010)
- Carroll (2007)
- Colburn (2000, 2003)
- Danhof (2005)
- De Gruttola (2001)
- Fine (2009)
- Fleming (1996)
- Fleming (2005)
- Frank (2003)
- Gollob (2006)
- Hurko (2009)
- Katz (2004)
- Kluft (2004)
- Krishna (2008)
- Kuhlmann (2006)
- Kumar (2006)
- Lassere (2007, 2008)
- Lathia (2009)
- Lee (2007)
- Lonn (2001)
- Park (2004)
- Psaty (1999, 2008)
- Shi (2009)
- Temple (1999)

9. Evidence for alternate, competing or compensatory pathways that can:
    a. Produce outcome through pathways independent of intervention's effect

    b. Produce nontherapeutic outcomes through pathways dependent on intervention

    c. Interfere with intervention's pathways

10. Strength of evidence that mechanism is similar to other interventions with known clinical effects

    - BDWG (2001)

11. Adverse effect mechanisms

    - Antoine (2009)
    - Chetty (2010)
    - Hurko (2009)
    - Kuhlmann (2006)
    - Sistare (2010)

Appendix B References

Altar CA, Bounos D Amakye D, et al. A prototypical process for creating evidentiary standards for biomarkers and diagnostics. Clin Pharmacol Ther. 2008 Feb; 83(2): 368-71.

Altar CA. The Biomarkers Consortium: on the critical path of drug discovery. Clin Pharmacol Ther. 2008 Feb; 83(2): 361-4.

Alymani NA, Smith MD, Williams DJ, et al. Predictive biomarkers for personalised anti-cancer drug use: Discovery to clinical implementation. Eur J Cancer. 2010 Mar;46(5):869-79.

Antoine DJ, Mercer AE, Williams DP, et al. Mechanism-based bioanalysis and biomarkers for hepatic chemical stress. Xenobiotica. 2009;39:565-77.

Bast RCJ, Lilja H, Urban N, et al. Translational crossroads for biomarkers. Clin Cancer Res. 2005;11:6103-8.

Bhattacharya S, Mariani TJ. Array of hope: expression profiling identifies disease biomarkers and mechanism. Biochem Soc Trans. 2009;37:855-62.

Bhogal N, Balls M. Translation of new technologies: from basic research to drug discovery and development. Curr Drug Discov Technol. 2008;5:250-62.

Biomarker Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar; 69(3): 89-95.

Boffetta P. Biomarkers in cancer epidemiology: an integrative approach. Carcinogenesis. 2010;31:121-6.

Carden CP, Banerji U, Kaye SB, et al. From darkness to light with biomarkers in early clinical trials of cancer drugs. Clin Pharmacol Ther. 2009;85:131-3.

Carroll KJ. Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology. Pharm Stat. 2007;6:253-60.

Cavero I. Optimizing the preclinical/clinical interface: an Informa Life Sciences conference 12-13 December, 2006. Expert Opin Drug Saf. 2007;6:217-24.

Chau CH, Rixe O, McLeod H, et al. Validation of analytic methods for biomarkers used in drug development. Clin Cancer Res. 2008;14:5967-76.

Chetty RK, Ozer JS, Lanevschi A, et al. A systematic approach to preclinical and clinical safety biomarker qualification incorporating Bradford Hill’s principles of causality association. Clin Pharmacol Ther. 2010 Aug;88(2):260-2.

Clark DP. Ex vivo biomarkers: functional tools to guide targeted drug development and therapy. Expert Rev Mol Diagn. 2009;9:787-94.

Coate LE, John T, Tsao MS, et al. Molecular predictive and prognostic markers in non-small-cell lung cancer. Lancet Oncol. 2009;10:1001-10.

Colburn WA. Biomarkers in drug discovery and development: from target identification through drug marketing. J Clin Pharmacol. 2003;43:329-41.

Colburn WA. Optimizing the use of biomarkers, surrogate endpoints, and clinical endpoints for more efficient drug development. J Clin Pharmacol. 2000;40:1419-27.

Collins CD, Purohit S, Podolsky RH, et al. The application of genomic and proteomic technologies in predictive, preventive and personalized medicine. Vascul Pharmacol. 2006;45:258-67.

Danhof M, Alvan G, Dahl SG, et al. Mechanism-based pharmacokinetic-pharmacodynamic modeling-a new classification of biomarkers. Pharm Res. 2005;22:1432-7.

Danna EA, Nolan GP. Transcending the biomarker mindset: deciphering disease mechanisms at the single cell level. Curr Opin Chem Biol. 2006;10:20-7.

Day M, Balci F, Wan HI, et al. Cognitive endpoints as disease biomarkers: optimizing the congruency of preclinical models to the clinic. Curr Opin Investig Drugs. 2008;9:696-706.

Day M, Rutkowski JL, Feuerstein GZ. Translational medicine—a paradigm shift in modern drug discovery and development: the role of biomarkers. Adv Exp Med Biol. 2009;655:1-12.

De Gruttola VG, Clax P, DeMets DL, et al. Considerations in the evaluation of surrogate endpoints in clinical trials. Summary of a National Institutes of Health workshop. Control Clin Trials. 2001;22:485-502.

Desai AA, Hysi P, Garcia JG. Integrating genomic and clinical medicine: searching for susceptibility genes in complex lung diseases. Transl Res. 2008;151:181-93.

Desar IM, van Herpen CM, van Laarhoven HW et al. Beyond RECIST: molecular and functional imaging techniques for evaluation of response to targeted therapy. Cancer Treat Rev. 2009;35:309-21.

Dhani N, Siu LL. Clinical trials and biomarker development with molecularly targeted agents and radiotherapy. Cancer Metastasis Rev. 2008;27:339-49.

Dudley JT, Butte AJ. Identification of discriminating biomarkers for human disease using integrative network biology. Pac Symp Biocomput. 2009;27-38.

Duffy SW, Treasure FP. Potential surrogate endpoints in cancer research—some considerations and examples. Pharm Stat. 2009.

Fasolo A, Sessa C. Translational research in phase I trials. Clin Transl Oncol. 2009;11:580-8.

Fine BM, Amler L. Predictive biomarkers in the development of oncology drugs: a therapeutic industry perspective. Clin Pharmacol Ther. 2009;85:535-8.

Fleming TR. Surrogate endpoints and FDA's accelerated approval process. Health Aff (Millwood). 2005;24:67-78.

Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125:605-13.

Floyd E, McShane TM. Development and use of biomarkers in oncology drug development. Toxicol Pathol. 2004;32 Suppl 1:106-15.

Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nat Rev Drug Discov. 2003;2:566-80.

Gimmi CD. Current stumbling blocks in oncology drug development. Ernst Schering Res Found Workshop. 2007;135-49.

Glassman RH, Ratain MJ. Biomarkers in early cancer drug development: limited utility. Clin Pharmacol Ther. 2009;85:134-5.

Gollob JA, Bonomi P. Historic evidence and future directions in clinical trial therapy of solid tumors. Oncology (Williston Park). 2006;20:10-8.

Goodsaid F, Frueh F. Biomarker qualification pilot process at the U.S. Food and Drug Administration. AAPS J. 2007;9:E105-8.

Goodsaid F, Papaluca M. Evolution of biomarker qualification at the health authorities. Nat Biotechnol. 2010;28:441-3.

Goodsaid FM, Frueh FW, Mattes W. Strategic paths for biomarker qualification. Toxicology. 2008;245:219-23.

Goulart BH, Clark JW, Pien HH, et al. Trends in the use and role of biomarkers in phase I oncology trials. Clin Cancer Res. 2007;13:6719-26.

Hampel H, Mitchell A, Blennow K, et al. Core biological marker candidates of Alzheimer's disease—perspectives for diagnosis, prediction of outcome and reflection of biological activity. J Neural Transm. 2004;111:247-72.

Hong H, Goodsaid F, Shi L, et al. Molecular biomarkers: a US FDA effort. Biomark Med. 2010;4:215-25.

Hunter DJ, Losina E, Guermazi A, et al. A pathway and approach to biomarker validation and qualification for osteoarthritis clinical trials. Curr Drug Targets. 2010;11:536-45.

Hurko O. The uses of biomarkers in drug development. Ann N Y Acad Sci. 2009;1180:1-10.

Jacobs A. An FDA perspective on the nonclinical use of the X-Omics technologies and the safety of new drugs. Toxicol Lett. 2009;186:32-5.

Jain KK. Cancer biomarkers: current issues and future directions. Curr Opin Mol Ther. 2007;9:563-71.

Katz R. Biomarkers and surrogate markers: an FDA perspective. NeuroRx. 2004;1:189-195.

Kelloff GJ, Boone CW, Crowell JA, et al. Surrogate endpoint biomarkers for phase II cancer chemoprevention trials. J Cell Biochem Suppl. 1994;19:1-9.

Kelloff GJ, Sigman CC. New science-based endpoints to accelerate oncology drug development. Eur J Cancer. 2005;41:491-501.

Kensler TW, Davidson NE, Groopman JD, et al. Biomarkers and surrogacy: relevance to chemoprevention. IARC Sci Publ. 2001;154:27-47.

Kluft C. Principles of use of surrogate markers and endpoints. Maturitas. 2004;47:293-8.

Krishna R, Herman G, Wagner JA. Accelerating drug development using biomarkers: a case study with sitagliptin, a novel DPP4 inhibitor for type 2 diabetes. AAPS J. 2008;10:401-9.

Kuhlmann J, Wensing G. The applications of biomarkers in early clinical drug development to improve decision-making processes. Curr Clin Pharmacol. 2006;1:185-91.

Kumar S, Mohan A, Guleria R. Biomarkers in cancer screening, research and detection: present and future: a review. Biomarkers. 2006;11:385-405.

Kurian S, Grigoryev Y, Head S, et al. Applying genomics to organ transplantation medicine in both discovery and validation of biomarkers. Int Immunopharmacol. 2007;7:1948-60.

Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. Eur J Cancer. 2007;43:2559-79.

Lassere MN. The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. Stat Methods Med Res. 2008;17:303-40.

Lassere MN, Johnson KR, Boers M, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. J Rheumatol. 2007;34:607-15.

Lathia CD, Amakye D, Dai W, et al. The value, qualification, and regulatory use of surrogate end points in drug development. Clin Pharmacol Ther. 2009;86:32-43.

Lavallie ER, Dorner AJ, Burczynski ME. Use of ex vivo systems for biomarker discovery. Curr Opin Pharmacol. 2008;8:647-53.

Lee JW, Figeys D, Vasilescu J. Biomarker assay translation from discovery to clinical studies in cancer drug development: quantification of emerging protein biomarkers. Adv Cancer Res. 2007;96:269-98.

Lesko LJ. Paving the critical path: how can clinical pharmacology help achieve the vision? Clin Pharmacol Ther. 2007;81:170-7.

Levinson SS. Weak associations between prognostic biomarkers and disease in preliminary studies illustrates the breach between statistical significance and diagnostic discrimination. Clin Chim Acta. 2010; 411:467-73

Lin D, Hollander Z, Meredith A, et al. Searching for ‘omic’ biomarkers. Can J Cardiol. 2009;25 Suppl A:9A-14A.

Lock EA, Bonventre JV. Biomarkers in translation; past, present and future. Toxicology. 2008;245:163-6.

Lockhart BP, Walther B. [Biomarkers: “Found in translation”]. Med Sci (Paris). 2009;25:423-30.

Lonn E. The use of surrogate endpoints in clinical trials: focus on clinical trials in cardiovascular diseases. Pharmacoepidemiol Drug Saf. 2001;10:497-508.

Malkesman O, Austin DR, Chen G, et al. Reverse translational strategies for developing animal models of bipolar disorder. Dis Model Mech. 2009;2:238-45.

Marrer E, Dieterle F. Biomarkers in oncology drug development: rescuers or troublemakers? Expert Opin Drug Metab Toxicol. 2008;4:1391-402.

Marrer E, Dieterle F. Promises of biomarkers in drug development—a reality check. Chem Biol Drug Des. 2007;69:381-94.

McMichael AJ, Hall AJ. The use of biological markers as predictive early-outcome measures in epidemiological research. IARC Sci Publ. 1997;281-9.

Merlo DF, Sormani MP, Bruzzi P. Molecular epidemiology: new rules for new tools? Mutat Res. 2006;600:3-11.

Moore RE, Kirwan J, Doherty MK et al. Biomarker discovery in animal health and disease: the application of post-genomic technologies. Biomark Insights. 2007;2:185-96.

Nicolette CA, Miller GA. The identification of clinically relevant markers and therapeutic targets. Drug Discov Today. 2003;8:31-8.

Oldenhuis CN, Oosting SF, Gietema JA, et al. Prognostic versus predictive value of biomarkers in oncology. Eur J Cancer. 2008;44:946-53.

Park JW, Kerbel RS, Kelloff GJ, et al. Rationale for biomarkers and surrogate end points in mechanism-driven oncology drug development. Clin Cancer Res. 2004;10:3885-96.

Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst. 2001;93:1054-61.

Potter JD. Cancer prevention: epidemiology and experiment. Cancer Lett. 1997;114:7-9.

Psaty BM, Lumley T. Surrogate end points and FDA approval: a tale of 2 lipid-altering drugs. JAMA. 2008;299:1474-6.

Psaty BM, Weiss NS, Furberg CD, et al. Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. JAMA. 1999;282:786-90.

Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer. 2005;5:142-9.

Ransohoff DF. Evaluating discovery-based research: when biologic reasoning cannot work. Gastroenterology. 2004a;127:1028.

Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. J Clin Epidemiol. 2007;60:1205-19.

Ransohoff DF. Promises and limitations of biomarkers. Recent Results Cancer Res. 2009;181:55-9.

Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer. 2004b;4:309-14.

Ratain MJ, Glassman RH. Biomarkers in phase I oncology trials: signal, noise, or expensive distraction? Clin Cancer Res. 2007;13:6545-8.

Richter WS. Imaging biomarkers as surrogate endpoints for drug development. Eur J Nucl Med Mol Imaging. 2006;33 Suppl 1:6-10.

Ryan NS, Fox NC. Imaging biomarkers in Alzheimer’s disease. Ann N Y Acad Sci. 2009;1180:20-7.

Sawyers CL. The cancer biomarker problem. Nature. 2008;452:548-52.

Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. Int J Clin Oncol. 2009;14:102-11.

Sinha A, Singh C, Parmar D, et al. Proteomics in clinical interventions: achievements and limitations in biomarker development. Life Sci. 2007;80:1345-54.

Sistare FD, Dieterle F, Troth S, et al. Towards consensus practices to qualify safety biomarkers for use in early drug development. Nat Biotechnol. 2010;28:446-54.

Steele VE, Kelloff GJ. Development of cancer chemopreventive drugs based on mechanistic approaches. Mutat Res. 2005;591:16-23.

Strand KJ, Khalak H, Strovel JW, et al. Expression biomarkers for clinical efficacy and outcome prediction in cancer. Pharmacogenomics. 2006;7:105-15.

Tan DS, Thomas GV, Garrett MD, et al. Biomarker-driven early clinical trials in oncology: a paradigm shift in drug development. Cancer J. 2009;15:406-20.

Taube SE, Clark GM, Dancey JE, et al. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. J Natl Cancer Inst. 2009;101:1453-63.

Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? JAMA. 1999;282:790-5.

Trusheim MR, Berndt ER, Douglas FL. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. Nat Rev Drug Discov. 2007;6:287-93.

Vainio H. Promise of molecular epidemiology—epidemiologic reasoning, biological rationale and risk assessment. Scand J Work Environ Health. 1999;25:498-504.

van Gool AJ, Henry B, Sprengers ED. From biomarker strategies to biomarker activities and back. Drug Discov Today. 2010;15:121-6.

Venitz J. Using exposure-response and biomarkers to streamline early drug development. Ernst Schering Res Found Workshop. 2007;47-63.

Verma M, Srivastava S. New cancer biomarkers deriving from NCI early detection research. Recent Results Cancer Res. 2003;163:72-84; discussion 264-6.

Verma M, Wright GLJ, Hanash SM, et al. Proteomic approaches within the NCI early detection research network for the discovery and identification of cancer biomarkers. Ann N Y Acad Sci. 2001;945:103-15.

Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev. 2007;16:1954-65.

Vineis P, Porta M. Causal thinking, biomarkers, and mechanisms of carcinogenesis. J Clin Epidemiol. 1996;49:951-6.

Wagner JA. Biomarkers: principles, policies, and practice. Clin Pharmacol Ther. 2009;86:3-7.

Wagner JA. Overview of biomarkers and surrogate endpoints in drug development. Dis Markers. 2002;18:41-6.

Wagner JA. Strategic approach to fit-for-purpose biomarkers in drug development. Annu Rev Pharmacol Toxicol. 2008;48:631-51.

Wagner JA, Prince M, Wright EC, et al. The Biomarkers Consortium: practice and pitfalls of open-source precompetitive collaboration. Clin Pharmacol Ther. 2010;87:539-42.

Wagner JA, Williams SA, Webster CJ. Biomarkers and surrogate end points for fit-for-purpose development and regulatory evaluation of new drugs. Clin Pharmacol Ther. 2007;81:104-7.

Warnock DG, Peck CC. A roadmap for biomarker qualification. Nat Biotechnol. 2010;28:444-5.

Williams SA, Slavin DE, Wagner JA, et al. A cost-effectiveness approach to the qualification and acceptance of biomarkers. Nat Rev Drug Discov. 2006;5:897-902.

Woodcock J. Chutes and ladders on the critical path: comparative effectiveness, product value, and the use of biomarkers in drug development. Clin Pharmacol Ther. 2009;86:12-4.

Woodcock J, Woosley R. The FDA critical path initiative and its influence on new drug development. Annu Rev Med. 2008;59:1-12.

Zhao L, Jin W, Rader D, et al. A Translational Medicine perspective of the development of torcetrapib: Does the failure of torcetrapib development cast a shadow on future development of lipid modifying agents, HDL elevation strategies or CETP as a viable molecular target for atherosclerosis? A case study of the use of biomarkers and Translational Medicine in atherosclerosis drug discovery and development. Biochem Pharmacol. 2009;78:315-25.

Appendix C. Workshop Proceedings Biological Mechanisms in Evidence-Based Medicine

Johns Hopkins Evidence-Based Practice Center

PI: Steven N. Goodman

Summary of Workshop Proceedings: Monday, November 30, 2009, 8:30 a.m to 4:00 p.m., Johns Hopkins Bloomberg School of Public Health

The goal of this one-day workshop was to bring together an interdisciplinary, international group of translational scientists, philosophers, toxicologists, historians of science, epidemiologists and evidence-based medicine researchers to explore the translation of biological mechanistic knowledge to effects of interventions in humans.

In attendance:

Participants: Lindley Darden, Ph.D., University of Maryland (by phone); David Eddy, M.D., Ph.D., Archimedes, Inc.; Charles Flexner, M.D., John Hopkins School of Medicine; John Groopman, Ph.D., Johns Hopkins Bloomberg School of Public Health; Jeremy Howick, M.Sc.,Ph.D., Oxford University/London School of Economics; David Kass, M.D., John Hopkins School of Medicine; Peter Keating, Ph.D., Université du Québec à Montréal (by phone); Gary Kelloff, M.D., National Cancer Institute; Scott Kern, Ph.D., John Hopkins School of Medicine; Malcolm Macleod, M.D., Ph.D., University of Edinburgh
Project Investigators: Steven Goodman, M.D., M.H.S., Ph.D., Johns Hopkins Bloomberg School of Public Health; Harry Marks, Ph.D., Johns Hopkins University; Karen Robinson, Ph.D., John Hopkins School of Medicine; Jason Gerson, Ph.D., Johns Hopkins Bloomberg School of Public Health; Emily Evans, M.P.H., Johns Hopkins Bloomberg School of Public Health.
Invited Guests: Steve Fox, M.D., S.M., M.P.H., AHRQ Project Officer; Gary Persinger, National Pharmaceutical Council.

I. Roadmap and workshop goals

Presenter: Steve Goodman

Following welcome and introductions, Dr. Goodman discussed workshop agenda and goals. Workshop being held early in project timeline, meant to be “creative brainstorming” session that informs development of conceptual framework. Key points:

- Presented conceptual map for the project: football field representing scale of probability of therapeutic efficacy. End zones represented by “Preclinical research,” and “Proof.” “Preclinical/mechanistic research,” Early developmental research,” and “clinical trials” represent 1-10 percent, 11-30 percent, 31-90 percent of the field, respectively, of probabilistic advances toward proof. Key idea: even if preclinical/mechanistic research affords “only” a 10 percent probability of an intervention working, still provides significant evidential value, and is preferable to simply picking therapies, devices or molecules at random.

- Reviewed U.S. Preventive Services Task Force criteria and analytic framework, with particular attention to how the task force defines “linkages” and “fit” between various steps in a mechanistic process.
- Presented domains of Biological Mechanisms in Evidence-Based Medicine (BMEBM) framework, including: (1) strength of evidence for existence of a pathway, (2) strength of evidence for the existence of a pathway in humans, and (3) completeness of pathway from intervention to clinical endpoint.
- Raised key questions for the group to consider over the course of the workshop:
    - Does framework capture relevant dimensions of “mechanism” to assess the strength of supporting evidence? What is missing?
    - Should framework focus on causal pathways? Is there a better conceptualization? Would different terminology be preferable?
    - How operational is framework, or any framework that tries to capture gaps in knowledge? To what extent can lack of knowledge be known in real time?
    - How should dimensions be rated/coded?
    - How reproducible are judgments likely to be?
    - How should dimensions be combined?

II. Invited Presentations

The Project Investigators invited subset of participants to give a brief presentation. Presentations intended to highlight their research and address ways in which research confronts questions related to biological mechanisms and preclinical evidence.

A. Summarizing the evidence from animal models of neurological disease: publication bias, poor internal validity, and (perhaps) some efficacy

Presenter: Malcolm Macleod

Dr. Macleod presented research concerning systematic reviews of animal models of interventions in stroke. Described search strategy/methods; presented findings from one review. Key points:

- Described how internal validity and external validity was assessed in systematic review.
- Suggested that reported efficacy (32 percent) of animal studies in stroke is reduced by publication bias (8 percent), randomization bias (6 percent), and comorbidity bias (14 percent), effectively reducing efficacy to 4 percent.
- Presented way of graphically summarizing evidence from animal studies. Graphical display indicates whether hypotheses have been confirmed or refuted, size of particular experiment, overall number of experiments in that field, and assessment of quality of evidence. Goal would be to use display to illustrate strength of evidence for each causal link hypothesized in a mechanism.

In discussion:

Eddy, Kass and Howick: clarification re: probability of finding efficacious intervention in humans, given that animal study was of “high quality,” according to Macleod’s definitions. Macleod: In 15 to 20 interventions his group has reviewed, only tPA demonstrated efficacy in humans. Virtually all animal experiments demonstrate an effect in animals, so that cannot be a barometer of the quality of animal studies.

Groopman: Difficulties of getting accurate pathology diagnosis in mice and how that contributes to measurement error.

Macleod: Scoring animal neurobehavior is difficult; there are challenges in measuring murine stroke effects; suggested ways to mitigate those errors.

B. Biological Mechanisms: A Perspective from Philosophy of Science

Presenter: Lindley Darden

Dr. Darden provided overview of key definitions and features of MDC (Machamer, Darden, Craver) account of biological mechanisms. Key points:

- Researchers interested in understanding biological mechanisms need to specify mechanisms in terms of entities and activities, start and set up conditions, finish and termination conditions, productive continuity, organization, degree of regularity (deterministic or probabilistic), character of the phenomena, spatial features, temporal features, and contextual features (e.g., integration of levels).
- MDC view distinguishes between mechanism *schema* (signaling more complete understanding; glass boxes) and mechanism *sketches* (signaling incompleteness; black and grey boxes). Dimensions of schemas include completeness, detail (abstract/specific), evidentiary support (how-possibly, how-plausibly, how-actually), and scope (generality).
- Reasoning strategies in examining mechanisms: construction, evaluation, and revision. Focused on evaluation, which seeks to remove incompleteness. Described moves from how- possibly/plausibly/actually: accumulation of experimental evidentiary support.
- Need to consider what kind of mechanism is being examined: (a) normal biological mechanism, (b) mechanism in disease, or (c) mechanism of intervening in the disease.

C. Use of biological mechanisms in the Archimedes model

Presenter: David Eddy

Dr. Eddy addressed the question: How can we use information on intermediate outcomes (e.g., mechanisms, biomarkers, surrogates) to draw conclusions about effects of treatments on health outcomes? He described the Archimedes model (AM). Key points:

- AM is built up from physiological pathways (at “clinical” or organ level, not cellular/molecular/genetic level), and organized around four submodels: individuals and population models, physiology models, health care delivery system model, and outcomes. Tests and treatments incorporated into an AM meant to reproduce disease mechanisms and diagnostic/health care processes. Presented model of coronary artery disease and MI occurrence to illustrate features and complexity of AM.
- Primary challenge for AM similar to BMEBM—determining whether mechanism is accurate predictor of health outcome. For AM, goal is to understand the relationship between intermediate variables and health outcomes. But many unknown variables—re: pathophysiological pathways and effects of intervention—can produce effects that:

  (1) are the reverse of what is expected, (2) in the expected direction but of the wrong magnitude, or (3) are as expected within some range but unexpected outside that range. Both AM and BMEBM trying to understand extent to which biomarkers, surrogate endpoints, or mechanisms produce unexpected effects.

- BMEBM should develop metrics to quantitatively measure gaps in knowledge of mechanisms: help predict “surprises” re: magnitude and direction of effects. Suggested doing survey of randomly chosen clinical trials of new treatments and count proportion in which effect on biomarkers as expected, but effect on health outcomes unexpected (i.e., by current medical theory or mathematical models that try to account for mechanisms).

In discussion:

Goodman: BMEBM will review surrogate outcomes literature—empirical attempt to represent the “probability of surprise,” since surrogates are, in a sense, mechanistic outcomes.

Macleod and Kern: Archimedes should be cautious re: how it incorporates findings from the clinical trial literature into models, since many types of bias and poor study quality could result in misleading conclusions.

Kass: Do “black boxes” that link variables represent statistical correlations?

Eddy: While an AM tries to capture all known causal relationships in particular disease or pathway, always something in the pathway about which we don’t know.

Kass: Given the sheer complexity of models, has AM simulation ever reached very counter-intuitive conclusions?

Eddy: Overall (qualitatively) this has not been the case; quantitatively, however, they have observed, at the level of specific variables, surprises in direction and magnitude.

D. Translating Basic Science to Prevention in High Risk Populations

Presenter: John Groopman

Dr. Groopman’s presentation addressed topic of risk/hazard assessment in carcinogenesis, with particular focus on formaldehyde and human cancers. Spoke of how International Agency for Research on Cancer (IARC) considers evidence from animal experiments in making determinations about what agents are carcinogenic to humans. Key points:

- Epidemiology driving evidence-gathering about risks in humans is decreasing, as occupational sites of potentially high exposure (e.g., manufacturing plants) have moved to developing world. More research in developed world focusing experimental model mechanistic data.
- If compound that has limited epidemiological data found to act through a “relevant mechanism,” (e.g., resulting in DNA damage), can be elevated from IARC Group 2A to Group 1.
    - Goodman: Is such a change in classification “merely” the function of arguments or a function of strong empirical evidence?
    - Groopman: Large number of compounds for which there will never be enough statistical power to do an epidemiological investigation. Absent epidemiological evidence, IARC will rely on experimental data emerging from mechanistic studies that assess DNA damage following exposure.
- Summarized IARC assessment of formaldehyde exposure risks. While overall assessment that formaldehyde belonged in Group 1 not contentious, much debate among IARC members about whether formaldehyde—a well-established cause of nasopharyngeal cancer—could also be labeled as cause of myeloid leukemia. Presented findings from recent epidemiological study of embalmers, formaldehyde exposure and cancer risk.

    - Goodman: Study might be a good example of a subgroup analysis that teases out emerging evidence about a hypothesized biological mechanism, one that may be based on prior mechanistic knowledge.
    - Marks: Formaldehyde may be a “reverse finding” story, in which the epidemiological finding subsequently generates research on the mechanism.

E. Reflections on the pathobiology of cardiac resynchronization therapy

Presenter: David Kass

Dr. Kass’s talk described lessons learned from development of cardiac resynchronization therapy (CRT). Described CRT and its effects on the heart, as well as measures used to assess whether CRT results in improvements in pump function and stroke volume. Key points:

- CRT was first cardiology intervention approved without mortality data. FDA approval of CRT was based on MIRACLE trial of 600 patients (relatively small sample), and not based on mortality, both anomalous in cardiology. CRT also approved before studies demonstrating *who* would clinically benefit were conducted. Basic research regarding mechanism was done post-approval; only now learning what mechanisms might be.
    - Goodman: Reason that CRT trial was so small and accepted surrogate endpoints rather than mortality was due to fact that CRT grounded in straightforward physiological model that we believe in and is fairly deterministic.
- Findings from clinical trials aimed at explaining why about 35 percent of patients with dyssynchrony who receive CRT—who seem like good candidates for the device—do not benefit from it. Largest study—PROSPECT trial—found that, for about half the patients, CRT offered no benefit (or whose dyssynchrony was worsened); in other half, patients who experienced perfect resynchronization still had highly variable surrogate endpoint improvements, suggesting no correlation between extent of resynchronization and long-term outcomes.
- Key messages from PROSPECT: (1) we either have inadequate measures of synchrony, or (2) reverse remodeling involves more than resynchronization (i.e., the proposed mechanism was wrong). As a result, Kass and colleagues have gone “back to the bench.” Through these experiments they’ve learned much about potential molecular mechanisms that may be involved in heart failure.

In discussion:

Flexner: Heart replacement was never subjected to a prospective randomized trial, but widely accepted as an effective therapy for heart failure. Why is evidence for these biomechanical procedures accepted with few qualifications, when more complex procedures seem to have a higher evidence threshold?

Kass: Perhaps it’s because heart replacement was based on a simple hemo-dynamic mechanism.

Kern and Goodman: There have been other seemingly simple mechanisms and procedures that we got wrong.

Kern: Part of what happens in development of mechanical interventions is that researchers take a narrow view of the problem; not concerned with off-target effects and assume there are no covariates about which to be concerned. When we test drugs we are always concerned with such things.

Kass: We've historically oversimplified heart failure story, and therapies we've developed either do not work, or, when they do seem to work (e.g., beta blockers and ACE inhibitors), we don't fully understand how they do.

III. Case study: Cancer targeted therapies: Gleevec

Discussant: Scott Kern

In order to ground discussion in a particular example, Project Investigators decided to include a case study concerning the development of the drug Gleevec to treat chronic myeloid leukemia (CML). Given his expertise in bench research broadly, and familiarity with development of Gleevec in particular, Investigators invited Dr. Kern to lead the discussion. Highlights of case study discussion:

- Kern discussed some of his work in developing a measure of pharmacogenetic synergy, a quantitative approach to capturing interactions between two agents or an agent and a disease. Discussed key concept in this work—pharmacogenetic window—which measures magnitude of the advantage achieved by a genetic classification of a subject.
- Kern: In some respects we were lucky with Gleevec because CML tumors have single, genetically simple tumor that the drug targeted. Key message of Kern's comments: in early stage drug development we ought to be using quantitative measures or a numerical threshold based on sound math to make decisions. He cautioned against relying on "stories"—namely scientists' discretion and qualitative judgments—in making determinations about potential of an agent.
- Marks: Tension between our incomplete understanding of causes of disease condition heterogeneity in humans on one hand and early drug development experiments (in vitro, animal models), which are intentionally designed to reduce or suppress heterogeneity, on the other. Kern: We should only accept evidence from simple, early developmental models as generalizable after we have introduced many other variables and see that our variable of interest still explains vast majority of overall variation.

IV. Reviewing the initial draft of the BMEBM conceptual framework

Dr. Goodman led discussion of domains contained in the initial draft of BMEBM conceptual framework. Highlights of discussion:

- Re: "strength of evidence for existence of intervention's pathway" domain. As presently conceived, this domain includes quality (design and execution) and strength (quantitative effect) of experimental evidence in preclinical models, as well as number and variety of experimental models. Goodman: We are trying to capture something about "robustness and the ability to predict," a domain that captures the extent to which preclinical studies have tried to reflect biological complexities.
    - Macleod: Transgenic models—now frequently used in pathway work that feeds into drug development—are part of an effort to better reflect this biological complexity.
    - Kelloff: Transgenic models, though by no means perfect, are better than in vitro cancer cell lines and leukemia transplant models used in the past. Knock-in experiments presently conducted in adult transgenic animals put in a specific gene thought to be involved in a pathway. Disadvantage of these experiments is that they are narrowed to one gene; advantage is that they match up the drugs that are molecularly targeted, and are able to test efficacy in an in vivo system.

- Re: “strength of evidence that the pathway exists in human disease states” domain.

  Kern: Closely consider what is meant by “strength of evidence.” Key question is: “Does a pathway variable that you are trying to change with a drug remain a dominant variable as you move closer to the natural heterogeneity of patients?”
- There was some discussion about the language contained in the framework. Does domain 1a refer to something like “internal validity?” Does domain 1b mean something like “generalizability or “repeatability?” Does domain 1c mean something like “robustness” or “repeatability across different models?”
- Kass: Framework should somehow capture the reality that the vast majority of the animal models that we currently use are not very predictive of human disease. With 90 percent of all basic research is done in small rodents, there is a significant translatability problem of animal models. Absence of genetic heterogeneity in animal models is further exacerbated by “co-pharmacotherapy” issue: no animal studies attempt to mimic human experience of exposure to other therapies meant to treat a particular disease.
- Eddy, Macleod, Goodman and Marks: Spoke about how an empirical component for BMEBM could be designed, in particular talking through issues of how a score/scale might be developed and challenges of data collection.
- Goodman: While these challenges are real, we must start from first principles, which is how the other hierarchies (e.g., GRADE) were built. One must begin by developing a common vocabulary for “what’s going to matter.” At present, we do not have any way to talk about these things.

The workshop adjourned.